

**Characterizing the tails of degree distributions in
real-world networks**

by

A. D. Broido

B.A., Boston College, 2012

M.S., University of Colorado, 2014

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Applied Mathematics

2019

This thesis entitled:
Characterizing the tails of degree distributions in real-world networks
written by A. D. Broido
has been approved for the Department of Applied Mathematics

Prof. Aaron Clauset

Prof. Jem Corcoran

Prof. Manuel Lladser

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Broido, A. D. (Ph.D., Applied Mathematics)

Characterizing the tails of degree distributions in real-world networks

Thesis directed by Prof. Aaron Clauset

This is a thesis about how to characterize the statistical structure of the tails of degree distributions of real-world networks. The primary contribution is a statistical test of the prevalence of scale-free structure in real-world networks. A central claim in modern network science is that real-world networks are typically "scale free," meaning that the fraction of nodes with degree k follows a power law, decaying like $k^{-\alpha}$, often with $2 < \alpha < 3$. However, empirical evidence for this belief derives from a relatively small number of real-world networks. In the first section, we test the universality of scale-free structure by applying state-of-the-art statistical tools to a large corpus of nearly 1000 network data sets drawn from social, biological, technological, and informational sources. We fit the power-law model to each degree distribution, test its statistical plausibility, and compare it via a likelihood ratio test to alternative, non-scale-free models, e.g., the log-normal. Across domains, we find that scale-free networks are rare, with only 4% exhibiting the strongest-possible evidence of scale-free structure and 52% exhibiting the weakest-possible evidence. Furthermore, evidence of scale-free structure is not uniformly distributed across sources: social networks are at best weakly scale free, while a handful of technological and biological networks can be called strongly scale free. These results undermine the universality of scale-free networks and reveal that real-world networks exhibit a rich structural diversity that will likely require new ideas and mechanisms to explain. A core methodological component of addressing the ubiquity of scale-free structure in real-world networks is an ability to fit a power law to the degree distribution. In the second section, we numerically evaluate and compare, using both synthetic data with known structure and real-world data with unknown structure, two statistically principled methods for estimating the tail parameters for power-law distributions, showing that in practice, a method based on extreme value theory and a sophisticated bootstrap and the more commonly used method

based an empirical minimization approach exhibit similar accuracy.

Dedication

To my family.

Acknowledgements

I would like to thank my committee members, Jem Corcoran, Dan Larremore, Manuel Llasder, Juan Restrepo, and Aaron Clauset for their support and guidance throughout graduate school, and in particular with compiling my thesis work. I would especially like to thank Aaron for taking on a student who was pretty far along in grad school already and had a very round-about trajectory through the program. Without his mentorship, patience, and understanding I have no doubt I wouldn't have made it.

The Clauset Lab has been a source of support and inspiration to deepen my research. I also want to thank the IQ Biology program, especially Amber McDonnell, Kristin Powell, and Andrea Stith for their continuous personal and administrative support. I'm lucky to have met so many wonderful people at CU who have been great friends to me throughout graduate school: Rachel Cox, Hillary Fairbanks, Abbie Jacobs, Taisa Kushner, Ryan Langendorf, Inom Mirzaev, John Nardini, Pat Sprenger, Nikki Sanderson, and Sam Way, this was so much better because of all of you.

I am so grateful to have a rich and diverse support network outside of my graduate program as well. Old friends Camilla Elvis, Leila Gerstein, Alessandra Sangurima, Corey Simpson, Jeremy Williams, and Rose Woodbury have stood by me for years and have continued to support me from afar. It's hard to put into words my feelings about Corey but she was without a doubt one of the most supportive friends I've ever had and I'm grateful for the 17 years we had together. In Colorado I made new friends through food and health: Kristin Burkholder, Jess Christin, Caitlin Gordon, Angela Huang, Shawn Mynar, and Hannah Schuette. I am especially grateful to Shawn for being a constant support and confidante, and for being my ally in the trenches during the worst

times for both of our health. Hannah is one of those rare people I connected with instantly and I'm so grateful to her for being my cheerleader and my study partner, and for keeping me sane this last year.

Finally, I want to thank my family. My parents have supported me in every way possible in whatever I've tried to do. It's really special to have such an open and close relationship with them and I'm immensely grateful for it. Lastly, Eric, who has been by my side as my friend, my colleague, and my partner through endless struggles, and has still managed to fill so much of our time with laughter and love. We did it.

Contents

Chapter

1	Introduction	1
1.1	Outline	4
2	Background	7
2.1	Power laws	7
2.1.1	Fitting the tail of a continuous power law	8
2.1.2	Fitting the tail of a discrete power law	9
2.2	Networks	10
2.2.1	Types of networks	11
2.3	Network models	14
2.3.1	Erdős-Rényi random graphs	15
2.3.2	Preferential attachment mechanism	16
2.3.3	Vertex-copying model	18
2.3.4	Configuration model	20
3	Prevalence of power laws in networks	22
3.1	Data analysis pipeline	23
3.2	Extracting degree distributions from real-world networks	25
3.3	Power-law analysis	30
3.3.1	Modeling power-law degree distributions	30

3.3.2	Testing goodness-of-fit	31
3.4	Alternative Distributions	32
3.4.1	Exponential	32
3.4.2	Log-normal	32
3.4.3	Power-law with exponential cutoff	34
3.4.4	Weibull (Stretched exponential)	34
3.4.5	Likelihood-ratio tests	35
3.5	Definitions of a scale-free network.	38
3.5.1	Parameters for defining scale-free networks	40
3.6	Empirical results	41
3.6.1	Scaling parameters.	41
3.6.2	Alternative distributions.	42
3.6.3	Assessing the scale-free hypothesis.	44
4	Robustness of results for power-law patterns in networks	49
4.1	Robustness checks	50
4.1.1	Results for simple networks alone	50
4.1.2	Results after removing power law with exponential cutoff from alternatives	51
4.1.3	Results after removing percent constraints	52
4.1.4	Results for the largest connected components alone	53
4.1.5	Results for scaling behavior of degree heterogeneity	54
4.1.6	Results of model comparisons using information criteria	55
4.2	Evaluating the method on synthetic data with ground truth	60
4.3	Discussion	63
5	Comparing methods for power-law fitting	67
5.1	Introduction	67
5.2	Methods	68

5.2.1	Kolmogorov Smirnov Method	69
5.2.2	Bootstrapping procedure	69
5.3	Results	74
5.3.1	Continuous synthetic data	74
5.3.2	Real Data	84
5.4	Discussion	89
6	Conclusion and Future Work	91
6.1	Conclusion	91
6.2	Future Work	92
	Bibliography	95

Tables

Table

3.1	Number of network data sets, and proportion of our network corpus, in each of five domains, under the taxonomy given by the <i>Index of Complex Networks</i> [23].	25
3.2	Comparison of scale-free and alternative distributions. The percentage of network data sets that favor the power-law model M_{PL} , alternative model M_{Alt} , or neither, under a likelihood-ratio test, along with the form of the alternative distribution $f(x)$	44
4.1	Comparison of scale-free and alternative distributions, using AIC. The percentage of network data sets that favor the power-law model M_{PL} , alternative model M_{Alt} , or neither, under a standard AIC comparison (see text), along with the form of the alternative distribution $f(x)$	57
5.1	Output of each method on continuous real data sets.	88

Figures

Figure

2.1	A simple graph. Reproduced from A. Clauset (2017) <i>Lecture 1: Network Basics</i> [Lecture notes]. University of Colorado Boulder, CSCI 5352	12
2.2	A bipartite graph and its one-mode projections. Reproduced from A. Clauset (2017) <i>Lecture 1: Network Basics</i> [Lecture notes]. University of Colorado Boulder, CSCI 5352	14
2.3	A weighted directed multigraph. Reproduced from A. Clauset (2017) <i>Lecture 1: Network Basics</i> [Lecture notes]. University of Colorado Boulder, CSCI 5352	15
2.4	Edős-Rényi random graphs at different choices of n and c . Reproduced from A. Clauset (2017) <i>Lecture 3: Random graphs I: homogeneous degrees</i> [Lecture notes]. University of Colorado Boulder, CSCI 5352	16
2.5	Networks grown by linear preferential attachment with $r = c = 1$ and $n = 5, 50$, and 1000. The choice of $c = 1$ means that each node has an out-degree of 1. Reproduced from A. Clauset (2017) <i>Lecture 12: Growing Networks</i> [Lecture notes]. University of Colorado Boulder, CSCI 5352	17
2.6	Networks grown by the vertex-copying mechanism with $q = 1/2$ and $n = 10, 100$, and 1000. The blue lines are the empirical degree distributions. Reproduced from A. Clauset (2017) <i>Lecture 12: Growing Networks</i> [Lecture notes]. University of Colorado Boulder, CSCI 5352	19

3.1	Mean degree $\langle k \rangle$ as a function of the number of nodes n . The 928 network data sets in the corpus studied here vary broadly size and density. For data sets with more than one degree sequence (see text), we plot the median of the corresponding set of mean degrees.	24
3.2	A graph simplification function, which takes as input a network G . In this case, if G is directed, the function returns three degree sequences: the in-degrees, out-degrees, and undirected degrees, while if G is undirected, it returns the degree sequence. 3.1 contains complete details.	26
3.3	Flowchart describing the path from network data set to degree sequence(s). Each step removes a layer from the properties. The gray path is for multiplex, bipartite, or multigraph networks, while the blue is for weighted networks without these properties. Details in text.	28
3.4	Taxonomy of scale-free network definitions. <i>Super-Weak</i> meaning that a power law is not necessarily a statistically plausible model of a network's degree distribution but it is less implausible than alternatives; <i>Weakest</i> , meaning a degree distribution that is plausibly power-law distributed; <i>Weak</i> , adds a requirement that the distribution's scale-free portion cover at least 50 nodes; <i>Strong</i> , adds a requirement that $2 < \hat{\alpha} < 3$ and the <i>Super-Weak</i> constraints; and, <i>Strongest</i> , meaning that every associated simple graph can meet the <i>Strong</i> constraints. The <i>Super-Weak</i> overlaps with the Weak definitions and contains the Strong definitions as special cases. Networks that fail to meet any of these criteria are deemed Not Scale Free.	39
3.5	Distribution of $\hat{\alpha}$ by scale-free evidence category. For networks with more than one degree sequence, the median estimate is used, and for visual clarity the 8% of networks with a median $\hat{\alpha} \geq 6.5$ are omitted.	42
3.6	Median $\hat{\alpha}$ parameter versus network size n . A horizontal band highlights the canonical $\alpha \in (2, 3)$ range and illustrates the broad diversity of estimated power-law parameters across empirical networks.	43

3.7	Proportion of networks by scale-free evidence category. Bars separate the Super-Weak category from the nested definitions, and from the Not Scale Free category, defined as networks that are neither Weakest or Super-Weak.	45
3.8	Proportion of networks by scale-free evidence category and by domain. (a) Biological networks, (b) social networks, and (c) technological networks. Tickers show change in percent from the pattern in all of the data sets.	47
4.1	Proportions of networks in each scale-free evidence category for simple networks. . .	51
4.2	Proportions of networks in each scale-free evidence category with removed degree percentage requirements.	52
4.3	Proportions of networks in each scale-free evidence category for directed networks with removed degree percentage requirements.	53
4.4	Moment ratio scaling. For 3662 degree sequences, the empirical ratio of the second to first moments $\langle k^2 \rangle / \langle k \rangle^2$ as a function of network size n , showing substantial variation across networks and domains, little evidence of the divergence pattern expected for scale-free distributions, and perhaps a roughly sublinear scaling relationship (smoothed mean via exponential kernel, with smoothed standard deviations).	55
4.5	Scatterplot of the degree heterogeneity factor for empirical and synthetic simple networks vs their size. Blue points are empirical networks and 12 synthetic networks were generated from the best power-law fit for each, shown in grey.	56
4.6	Proportions of networks in each scale-free evidence category using AIC instead of LRT for comparison of alternative distributions. Tickers indicate percent change from the results in Chapter 3.	57
4.7	Proportions of simple networks in each scale-free evidence category using AIC instead of LRT for comparison of alternative distributions. Tickers indicate percent change from the results for simple networks in Chapter 3.	58

5.1	Distribution that is exponential below x_{\min} and power law above. (a) Distribution with discontinuous derivative and various x_{\min} values. Specifically, $b = 2$ (see text for details). (b) Distribution with continuous derivative (see text for details).	75
5.2	Recovery of x_{\min} for synthetic data sets with discontinuous derivative. Each data set has 10000 points, 250 at each true x_{\min} value. The points are at the median \hat{x}_{\min} over all 250 estimates and the 25th to 75th quantile range is shaded.	77
5.3	RMSE convergence comparison for bootstrap method on synthetic data sets with discontinuous derivative by number of bootstrap resamples. 100 data sets each. The dashed line is the theoretical limit $n^{-1/2}$	78
5.4	RMSE convergence comparison for bootstrap method on synthetic data sets with discontinuous derivative by number of bootstrap resamples. 250 data sets each. The dashed line is a linear regression (fit to points with $\text{RMSE} < 0.1$) with slope -4.97×10^{-6} and intercept 0.0435.	78
5.5	Recovery of x_{\min} for synthetic data sets with continuous derivative. Each data set has 10000 points, 100 at each x_{\min} value. The median is plotted with the 25th to 75th quantile range shaded.	79
5.6	RMSE convergence comparison for the two methods on C^1 synthetic data sets by size with 500 bootstrap resamples. Each point is the RMSE over 250 synthetic data sets. The dashed line is the theoretical limit $n^{-1/2}$	80
5.7	RMSE convergence comparison for bootstrap method on C^1 synthetic data sets by number of bootstrap resamples.	81
5.8	Pareto distribution	81
5.9	Recovery of x_{\min} for pareto synthetic data sets.	82
5.10	Pareto deviate distribution	82
5.11	Recovery of x_{\min} for pareto deviate synthetic data sets. The estimate of α is shown above.	83

Chapter 1

Introduction¹

Networks are a powerful way to both represent and study the structure of different kinds of complex systems. Examples today are plentiful and include social interactions among individuals, both offline and online, protein or gene interactions in biological organisms, communication between digital computers, and various kinds of transportation systems. Across scientific domains and different types of networks, it is common to encounter the claim that most or all real-world networks are *scale free*. The precise details of this claim vary across the literature [6, 87, 63, 69, 2, 51, 111], but it is generally agreed that a network is scale free if the fraction of nodes with degree k follows a power-law distribution $k^{-\alpha}$, where $\alpha > 1$. Some versions of this “scale-free hypothesis” make the requirements stronger, e.g., requiring that $\alpha \in (2, 3)$ or that node degrees evolve by the preferential attachment mechanism [30, 11]. Other versions make them weaker, e.g., requiring that the power law holds only in the upper tail [109], can exhibit an exponential cutoff [84], or is merely more plausible than a thin-tailed distribution like an exponential or normal [7].

No matter the definition, people are very interested in describing degree distributions of empirical networks. In biology, it is common to hear some version of the phrase “structure determines function.” This idea carries over into network science as well: we can learn a lot about the functions of the processes described by a network if we understand the structure of the network. Degree distributions are one aspect of network structure that has undergone a great deal of study. In this thesis, we focus primarily on power-law distributions, and assessing the fit of power laws

¹ This chapter is adapted from: **A. D. Broido** and A. Clauset. Scale-free networks are rare. *Nature Communications* **10**: 1017 (2019)

to empirical degree sequences. The presence of power-law degree distributions overlaps with some definitions of scale-free networks, and this guides our analysis.

The study and use of scale-free networks is widespread throughout network science [6, 11, 19, 74, 70]. Many studies investigate how the presence of scale-free structure shapes the dynamics of processes running over a network [40, 51, 81, 111, 74, 83, 4, 3, 78, 73]. For example, in the Kuramoto model of oscillator dynamics, a transition to global synchronization is well-known to occur at a precise threshold K_c , whose value depends on the power-law parameter α of the degree distribution [58, 92, 50, 94, 93]. Scale-free networks are also widely used as a substrate for network-based numerical simulations and experiments, and the study of specific generating mechanisms for scale-free networks has been framed as providing a common basis for understanding network assembly [86, 97, 11, 30, 82, 14, 60, 63].

The universality of scale-free networks, however, remains controversial. Many studies find support for their ubiquity [69, 2, 40, 81, 38, 48, 12], while others challenge those claims on statistical or theoretical grounds [103, 61, 87, 63, 109, 101, 41, 102, 54, 56, 1, 31]. This conflict in perspective has persisted because past work has typically relied upon small, often domain-specific data sets, less rigorous statistical methods, differing definitions “scale-free” structure, and unclear standards of what counts as evidence for or against the scale-free hypothesis [90, 69, 2, 40, 51, 81, 111, 80, 96]. Additionally, relatively few studies have performed statistically rigorous comparisons of fitted power-law distributions to alternative, non-scale-free distributions, e.g., the log-normal or the stretched exponential, which can imitate a power-law form in realistic sample sizes [22]. These issues raise a natural question: just how pervasive is strong empirical evidence of scale-free structures in real-world networks of different kinds?

Central to this debate are the difficulties that arise from the diversity of uses of the term “scale-free network.” The classic definition [6, 61, 78, 12] states that a network is scale free if its degree distribution $\Pr(k)$ has a power law $k^{-\alpha}$ form. A power law is the only normalizable density function $f(k)$ for node degrees in a network that is invariant under rescaling, i.e., $f(c k) = g(c)f(k)$ for any constant c [74], and thus “free” of a natural scale. For the degree distribution of a network,

being scale free implies a power-law pattern, and vice versa. But, scale invariance can also refer to non-degree-based aspects of a network’s structure, e.g., a network’s subgraphs may be structurally self-similar [100, 10], and these networks are sometimes also called scale free.

Scale-free networks are commonly discussed in the literature on mechanisms for network assembly, and particularly in the context of the preferential attachment model [86, 97, 6], in which the probability that a node gains a new connection is proportional to its current degree k . Although preferential attachment is perhaps the most famous mechanism that produces scale-free networks, there exist other mechanisms that can produce scale-free networks without using preferential attachment [19, 74, 70]. And, some variations of preferential attachment do not produce networks with power-law degree distributions [12], although sometimes those networks are still, confusingly, called scale free. On the other hand, the shape of a degree distribution itself imposes only modest constraints on overall network structure [8] and thus represents relatively weak evidence for distinguishing generating mechanisms [71, 67, 79, 91], even when its underlying form is known. For heavy-tailed distributions, in particular, identifying that form from data is non-trivial, and log-normal distributions often fit empirical data as well or better than power laws [91, 89, 22].

For example, one recent study [35] used rigorous methods for fitting and testing for power-law distributions [22] to investigate the singular values of the adjacency matrix, the eigenvalues of the Laplacian, and the degree distributions of a number of real-world networks. Although this study claimed to find broad evidence of scale-free structure in these networks, the evidence remains ambiguous in two crucial ways. First, the statistical plausibility of scale-free structure is strongest in the singular and eigenvalue analyses rather than in the degree distributions, which is a different kind of scale-free structure than the hypothesis typically posits. Second, the analyses did not include controls for spurious conclusions due to small sample sizes or comparisons against alternative distributions, which serve to reduce the likelihood of false positives [22].

1.1 Outline

The primary goal of this thesis is to rigorously investigate the presence of power-law structure in networks and methods for this study. Here we outline the work we have done.

Chapter 2 introduces foundational concepts about power laws and networks. Later chapters analyze various different types of networks so we make sure to define those here. We also define degree sequences, and outline methods for fitting power-law distributions to degree sequences. This chapter defines the basic ideas and notation that are necessary to understand the analysis presented in subsequent chapters.

With the fundamentals out of the way, we can begin to discuss our work. Chapter 3 is the central work of this thesis, in which we develop a statistical test of the prevalence of scale-free structure in real-world networks. Testing empirical networks for scale-free structure requires empirical data sets. Previous work in this area performed analysis on relatively small corpora of simple networks. We use 928 networks from the *Index of Complex Networks* (ICON), which is a large and diverse index of real-world networks from a wide range of scientific disciplines [23]. Additionally, many of these networks are not simple. As a result, we introduce new methodology to process all the networks in our corpus, not just the simple ones, that involves extracting multiple degree sequences from each network and then combining the results for all of them to report results for the network as a whole.

We explain the methodology we use for fitting a power-law to the data and present the four alternative models that we assess as well. Chapter 3 also explains the likelihood ratio tests used to compare these alternative models to the power-law model. We then present the results for the corpus, describing the power-law fits, the outcomes of the likelihood ratio tests, and the proportions of data sets that fit into each of our scale-free definitions. Finally, we examine the difference in the distribution across scale-free categories when we look only at one scientific domain at a time. Overall we find that scale-free patterns appear a lower frequency in empirical networks than much of the literature would suggest.

Because our results do not align with the received view in some sectors of the network science community, we extensively assess their robustness with a series of tests. In Chapter 4 we present these tests and their results. One concern is that the methodology we introduce to include non-simple graphs by extracting multiple degree sequences from one network may bias the results in some unpredictable way. To account for this, we find the results for more permissive versions of our scale-free definitions that require that only one of the corresponding degree sequences for a given network satisfy the requirements for that definition. The results are qualitatively similar to the results in Chapter 3.

In Chapter 4 we also present an analysis of the scaling behavior of the ratio of the first and second moments of the degree sequences. Scale-free distributions have finite moments $\langle k^m \rangle$ only for $m < \alpha - 1$. If the distribution falls into the canonical range $2 < \alpha < 3$, this means the mean of the distribution is finite but the second moment is infinite. Thus in this range, the moment ratio $\langle k^2 \rangle / \langle k \rangle^2$ is divergent. The hypothesis motivating this test is that as network size grows, we expect to see a divergent trend in the moment ratio. Though of course our data sets are finite, the range of sizes present in our corpus should be enough to give an indication of this pattern. There is no obvious divergence trend in the data, and whatever pattern is present is unclear.

This chapter presents five additional tests, each of which slightly alters the definitions, the data considered, or the methodology. We also include a test of the results on synthetic data, where we know exactly how the data should be classified. The overall conclusions from the tests are consistent with the results from Chapter 3.

Central to all of the above analysis of power-law patterns in degree distributions is an ability to accurately fit power-law distributions to data. In particular, we fit the tail of a power law so we need to estimate the value x_{\min} where this tail begins. This is not a trivial task because fluctuations are greatest in the upper tail of power law data, meaning the data is sparsest where we need precision in estimates of the model parameter and threshold [24]. All of our analysis on networks uses the minimum distance spanning procedure (MDSP) to find the best x_{\min} value for a given data set. In Chapter 5 we compare this method to an alternative method that uses bootstrapping

to estimate x_{\min} . The estimate of α depends on the estimate of x_{\min} . This bootstrapping method asymptotically minimizes the mean squared error (AMSE) in the estimate of α . We compare the methods on different types of synthetic data, examining areas where each method excels or struggles. While discrete-valued data is more relevant for analysis of degree sequences, much of this chapter focuses on continuous distributions because the theory for the bootstrapping method is more obviously defined here. We also test both the MDSP and bootstrapping method on empirical data sets and here we include some discrete data sets in addition to continuous data sets. This analysis is exploratory and may inspire future study of the selection between methods for power-law fitting. Finally we conclude with Chapter 6, where we discuss implications of the results of previous chapters, and discuss future work.

Chapter 2

Background

In Chapter 1 we discussed the widespread use of network models and power-law models. To understand the implications and use of these models, it is important to understand the basic underlying ideas. Here we go over some definitions and notation, explaining what networks and power-laws are, and some useful properties they have.

2.1 Power laws

Many observations we make on a regular basis come from data sets that are well represented by the mean of the data set. For example, the average height of players on an NBA team gives a pretty good sense of how tall professional basketball players are. Even the players who are unusually short or unusually tall are far less than a factor of two different from the mean.

Some data sets are not well-described by their average, however, especially when extreme values are common. For example, the mean size of a city in the U.S. according to the 2000 Census was 9000 people, while the median is 1008 people. This gives us a good sense already that the mean is not a very good descriptor by itself. Instead, this type of data set is often better fit by a heavy-tailed distribution like a power law.

A power law is a probability distribution of the form

$$p(x) = Cx^{-\alpha}$$

where $\alpha > 1$ and $C > 0$ are constant. Often in empirical data, a power-law is a good description for the larger values only. We refer to this as the *tail* of the data and say that values above some

minimum threshold value x_{\min} follow a power law. While it may be easy to see that a data set has what's referred to as a "heavy tail", there are several distributions that fit that description. It can be tricky to show that a power-law distribution is the best fit. This explores several methods for fitting power-law distributions to the tail of data sets.

2.1.1 Fitting the tail of a continuous power law

Fitting a power-law distribution to the tail of a data set requires finding the values of α and x_{\min} that best describe the data. We very often estimate α by maximizing a likelihood function.

A continuous power-law distribution for values above some minimum value x_{\min} must satisfy

$$1 = \int_{x_{\min}}^{\infty} Cx^{-\alpha}.$$

Thus

$$\begin{aligned} 1/C &= \int_{x_{\min}}^{\infty} x^{-\alpha} \\ &= \frac{1}{1-\alpha} x^{1-\alpha} \Big|_{x_{\min}}^{\infty} \\ &= \frac{1}{\alpha-1} x_{\min}^{1-\alpha} \end{aligned}$$

so

$$C = (\alpha - 1)x_{\min}^{\alpha-1}.$$

Thus this tail power law has the form

$$f(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha}.$$

Given a continuous data set $\vec{X} = \{x_1, x_2, \dots, x_n\}$, where $x_i > x_{\min}$ for all $i \in 1, \dots, n$, the likelihood function is

$$\begin{aligned} l(\vec{X}) &= \prod_{i=1}^n f(x_i) \\ &= \prod_{i=1}^n \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha}. \end{aligned}$$

The log-likelihood function is then

$$\mathcal{L}(\vec{X}) = n \log(\alpha - 1) - n \log x_{\min} - \alpha \sum_{i=1}^n \log \frac{x_i}{x_{\min}}.$$

To find the *maximum likelihood estimate* (MLE) for α , we find the value of α that maximizes this function. The derivative of \mathcal{L} with respect to α is

$$\frac{\partial}{\partial \alpha} \mathcal{L}(\vec{X}) = \frac{n}{\alpha - 1} - \sum_{i=1}^n \log \frac{x_i}{x_{\min}}.$$

Setting that equal to zero and solving for α gives the MLE for α

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^n \log \frac{x_i}{x_{\min}} \right]^{-1}. \quad (2.1)$$

Note that this estimate depends on x_{\min} . For this reason we often estimate x_{\min} and α simultaneously in practice. We estimate x_{\min} using the minimum distance spanning procedure, which we explain in detail in Section 5.2.1. The basic idea is that we choose x_{\min} to minimize the distance between the empirical distribution of the data set and the power-law tail distribution evaluated for the data set

$$D = \max_{x \geq x_{\min}} |E(x) - P(x | \hat{\alpha})|. \quad (2.2)$$

This estimate requires a fixed α so we calculate the MLE for α at each potential x_{\min} value, and then choose the combination with the smallest value for Eqn. 2.2.

2.1.2 Fitting the tail of a discrete power law

Discrete data is quite common and is especially relevant to this thesis because we analyze integer-valued data. The discrete power law must satisfy

$$\begin{aligned} 1/C &= \sum_{x=x_{\min}}^{\infty} x^{-\alpha} \\ &= \sum_{x=0}^{\infty} (x + x_{\min})^{-\alpha}. \end{aligned}$$

The right-hand side of this last equality is known as the Hurwitz-zeta function, denoted $\zeta(\alpha, x_{\min})$.

Thus the discrete power-law distribution for tail data is

$$f(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})}.$$

The MLE for α for the discrete distribution is harder to find than in the continuous case.

The likelihood function is

$$l(\vec{X}) = \prod_{i=0}^n \frac{x_i^{-\alpha}}{\zeta(\alpha, x_{\min})}$$

so the log-likelihood is

$$\mathcal{L}(\vec{X}) = -\alpha \sum_{i=0}^n \log x_i - n \log \zeta(\alpha, x_{\min}).$$

If we take a derivative and set this equal to zero, we cannot solve analytically for α . Typically we use numerical optimization schemes to get an estimate for α .

The procedure for the estimate of x_{\min} is the same in both the discrete and continuous cases: at each value of x_{\min} we numerically estimate the MLE for α , then choose the combination that minimizes (Eq. 2.2). Note that x_{\min} is not a model parameter in the traditional sense because it controls how much of the data is used in the likelihood calculation. Hence x_{\min} cannot be estimated directly using maximum likelihood, because that function is maximized when x_{\min} is the largest value. This is one reason why working with tail models is non-trivial.

2.2 Networks

A *network*, sometimes referred to as a *graph*, is a collection of nodes (or vertices) connected by edges (or links). We typically denote a network as $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of edges. Graph theory is the branch of mathematics that deals with the study of these objects and their properties. Typically the term network is used in more applied settings when referring to real data, and this is the term we will use to refer to data sets of this type.

Networks can be used to model any system that can be represented as a set of objects and their pairwise interactions. For example [23]:

The simplest kind of graph is appropriately called a *simple graph*:

Definition 1. (Simple graph.) A *simple graph* is a graph G in which edges are symmetrical and unique. The edge (u, v) is the same as (v, u) , and denotes the connection between nodes u and v .

Network	Node	Edge
Star Wars	character	co-occurrence
cat brain	anatomical region	interaction
Medieval Russia river trade	town	economic trade
HIV transmission	person	transmission
US airports	airport	commercial flight
Wikipedia	article	hyperlink

E can include at most once instance of each pair (u, v) , and for all u , edges of the form (u, u) are not included.

Edges that form between a node and itself, which are not included in simple graph, are called *self-loops*. Figure 2.1 shows a small example of a simple graph. The size of each node scales with its degree.

Definition 2. (Degree.) The *degree* of a node u in a simple graph $G = (V, E)$ is the number of edges in E that contain u .

Every node in a network has a degree. Combining all of them in a list gives the *degree sequence* of a network. The probability distribution $p(k)$ describing the probability that a node selected uniformly at random has degree k is called the *degree distribution*. There has been a great deal of study about degree distributions of networks [22, 50, 41, 73, 11, 90, 71, 5, 62, 81], some of which we will address in the coming sections.

2.2.1 Types of networks

There exist many different types of networks, each with different properties. As we saw, a simple graph is the most basic, with no properties assigned to nodes and edges. All other network types have some attributes on nodes and edges. Here we present a few network properties that we'll use in later chapters.

Possibly the most common type of non-simple graph is a *directed graph*, in which edges have a direction.

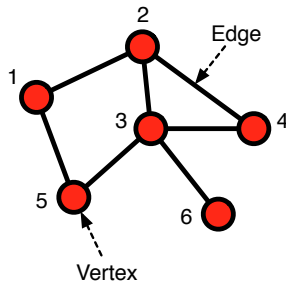


Figure 2.1: A simple graph. Reproduced from A. Clauset (2017) *Lecture 1: Network Basics* [Lecture notes]. University of Colorado Boulder, CSCI 5352

Definition 3. (Directed graph.) A *directed graph* is a graph $G = (V, E)$ in which edges can have direction. In a directed graph, the edges (u, v) and (v, u) are distinct entities.

To get a degree sequence for a network that is not simple, we have to make some choices about what kinds of edges to consider. A directed network has three associated degree sequences. The *in-degree sequence* considers the degree of a node to be the number of edges pointing into it. The *out-degree sequence* considers the degree of a node to be the number of edges pointing out of it. Finally, the degree sequence ignoring direction of edges just considers the total number of edges connecting into or out of a node.

A *path* between nodes u and y in a network can be thought of as a list of edges $[(u, v), \dots (x, y)]$ where each edge is in the form (source node, target node), and u is the source in the first edge and y is the target in the last edge. A path connects two nodes. A *cycle* is a path that connects a node back to itself. An *acyclic* graph is a graph that has no cycles. Directed acyclic graphs are sometimes referred to as DAGs.

Edges can also have a weight, indicating the strength of the relationship between two nodes. Weighted edges appear in a *weighted network*.

Definition 4. (Weighted graph.) A *weighted graph* is a graph $G = (V, E)$ in which edges can have weight. Here edges look like (u, v, w) where $w \in \mathbb{R}^+$ indicates the weight on the edge between nodes u and v .

Multigraphs also treat edges differently, allowing for duplicate edges between node pairs.

Definition 5. (Multigraph.) A *multigraph* is a graph $G = (V, E)$ in which edges are not unique so an edge (u, v) may appear in E more than once.

Multigraphs, weighted graphs, and directed graphs can all be *simplified*, which means we somehow eliminate the edge properties. A common and easy way to do this is to ignore any edge properties like weight or direction, and ignore the count on any edges with multiplicity greater than 1. Another way to simplify weighted networks specifically is to apply a threshold to the weights, keeping edges with weight above a certain threshold and discarding edges with weight below the threshold.

Sometimes graphs may have different edge types as in the case of a *multiplex network*.

Definition 6. (Multiplex graph.) A *multiplex graph* is a graph $G = (V, E)$ in which edges can have a type. Edges take the form (u, v, t) where t is a categorical variable representing the type of the connection between nodes u and v .

The edge types are sometimes referred to as *layers* of the network. For example, the network of US airports, where edges connect airports that have flights between them can be thought of as a multiplex or multilayer network, where the layers represent the set of edges flown by different airlines. Boston and Denver are connected by an edge in, say, the Southwest Airlines layer and another in the United Airlines layer. Labeling the edges allows us to separate the layers if we want to. In our airlines example, this would allow us to look at maps of where a specific airline flies, letting us answer a very different set of questions than those we could answer looking at the union of all the layers.

A temporal network also has edge types, but these are time stamps. For example, in an evolving social network where edges represent friendships and edge types indicate snapshots in time, edges will appear and disappear as friendships form and dissolve.

Nodes can also have types. In *bipartite networks*, nodes are assigned one of two types, and nodes of the same type are never connected.

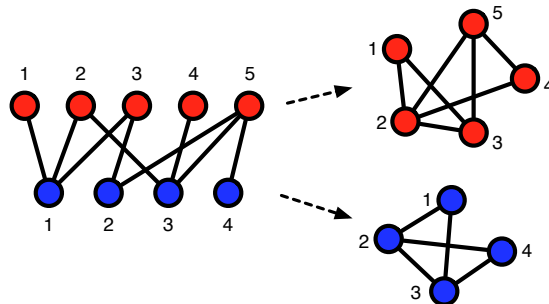


Figure 2.2: A bipartite graph and its one-mode projections. Reproduced from A. Clauset (2017) *Lecture 1: Network Basics* [Lecture notes]. University of Colorado Boulder, CSCI 5352

Definition 7. (Bipartite graph.) A *bipartite graph* is a graph $G = (V_a, V_b, E)$ in which the vertex set is broken into two types: V_a and V_b . Edges form only between nodes in V_a and V_b , never within a set.

It is possible to have more than two node types. This is called a k -partite graph, where k is the number of node types.

Any of these networks with types on edges or nodes, namely multiplex, temporal, and bipartite networks, can be *projected* onto a specific type. For example, in our US airports example we discussed looking at individual airline layers. This is formally called projecting onto a specific airline. In a temporal network, a *projection* is a particular snapshot in time. In a bipartite network, a projection onto nodes of single type is called an *one-mode projection*, and consists of only the nodes of that type. Edges in this projection indicate that two nodes of this type were connected to the same node of the other type. Any resulting multi-edges are often simplified. Figure 2.2 shows a bipartite graph and its one-mode projections.

A network may have any combination of these properties. Figure 2.3 shows a weighted, directed multigraph.

2.3 Network models

In any system, if we can understand something about the process by which it was formed, we can use that structural insight to better understand the system itself. For example, understanding

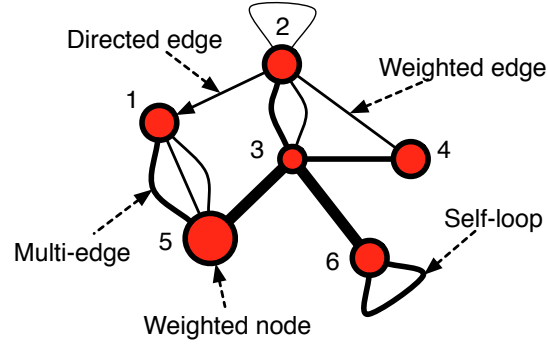


Figure 2.3: A weighted directed multigraph. Reproduced from A. Clauset (2017) *Lecture 1: Network Basics* [Lecture notes]. University of Colorado Boulder, CSCI 5352

how bonds form in a crystal structure can help scientists to synthesize new solids with similar structure. Similarly, understanding the generating mechanism of a network yields insights into the system we are studying. Network models have been studied a great deal in the literature [70, 91, 14, 15, 2, 60, 36, 73]. Here we present four network models that will be used in this thesis.

2.3.1 Erdős-Rényi random graphs

The original random graph model is the Erdős-Rényi model, named after Paul Erdős and Alfred Rényi. This model yields simple graphs. We denote the model $G(n, p)$, where n is the number of nodes and p is the probability of an edge forming between two nodes.

The Edős-Rényi model is not generally a good model for real-world networks. However, it is simple and easy to work with, and often provides a good null model for comparison. The degree distribution for $G(n, p)$ is a binomial distribution

$$f(k) = \binom{n-1}{p} p^k (1-p)^{n-1-k}$$

with $n-1$ independent trials, since a node isn't permitted to form an edge with itself. The mean degree for an Edős-Rényi graph is then given by $c = (n-1)p$. It is common to choose an ideal mean degree when generating a random graph. Since c is not a parameter of $G(n, p)$, we often set $p = c/(n-1)$, to enforce the choice of c . Note that for small p in the limit of large n , this distribution is well approximated by the Poisson distribution with mean and variance c . As small

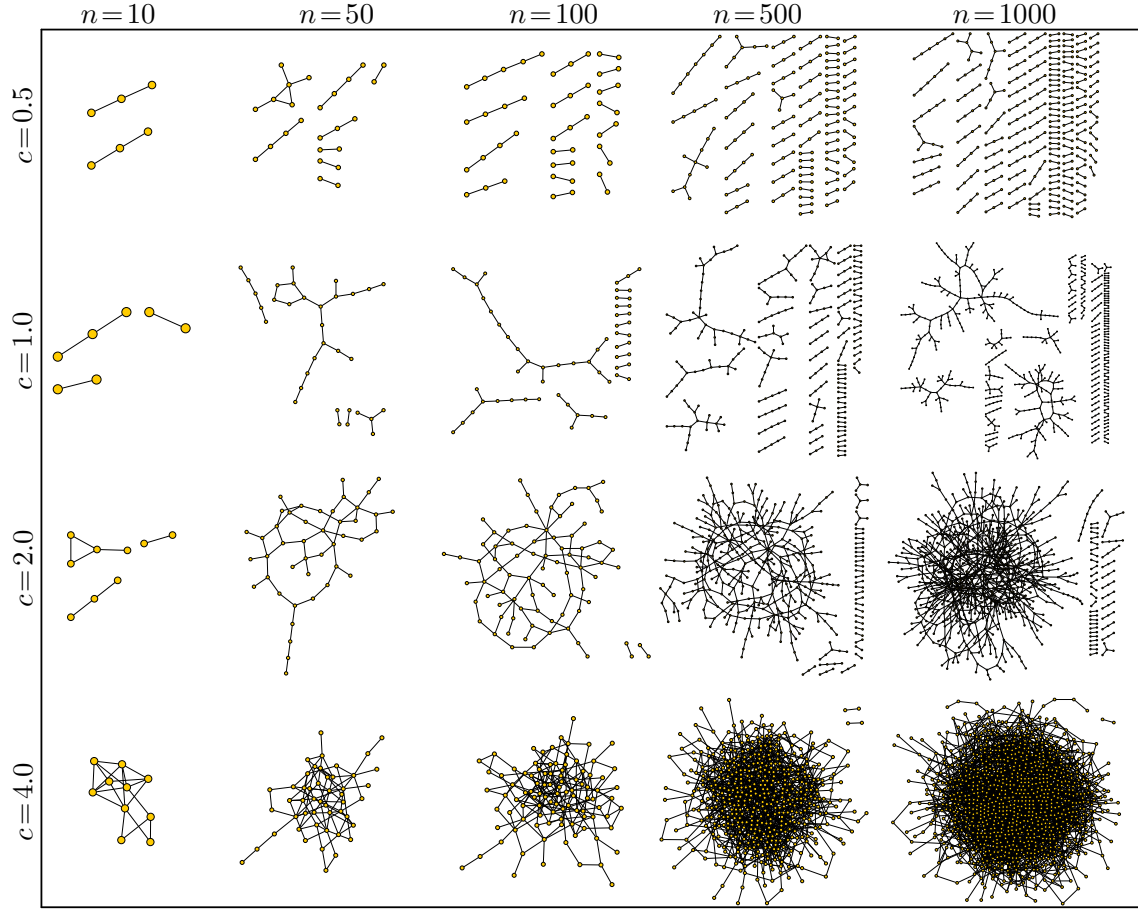


Figure 2.4: Edős-Rényi random graphs at different choices of n and c . Reproduced from A. Clauset (2017) *Lecture 3: Random graphs I: homogeneous degrees* [Lecture notes]. University of Colorado Boulder, CSCI 5352

p corresponds to sparse graphs and these are what is most commonly studied, we often say that Edős-Rényi networks have Poisson degree distributions.

Figure 2.4 shows instances of this model for different choices of n and c . Note that as c increases, the nodes become more connected.

2.3.2 Preferential attachment mechanism

The preferential attachment mechanism is one of the most famous network models, and it is the most relevant to this thesis. Like the Erdős-Rényi model, it describes how a network grows,

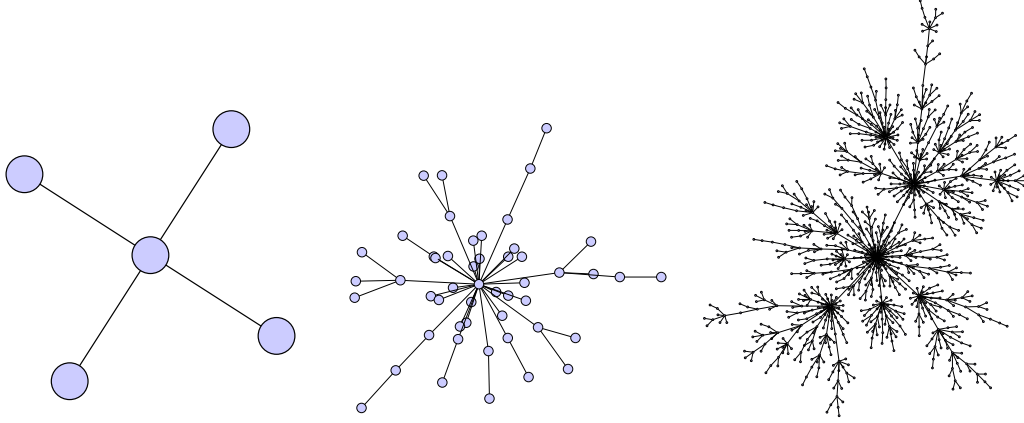


Figure 2.5: Networks grown by linear preferential attachment with $r = c = 1$ and $n = 5, 50$, and 1000 . The choice of $c = 1$ means that each node has an out-degree of 1. Reproduced from A. Clauset (2017) *Lecture 12: Growing Networks* [Lecture notes]. University of Colorado Boulder, CSCI 5352

adding new nodes and edges. This model has several different versions.

2.3.2.1 Linear preferential attachment

The Price model, named after Derek de Solla Price, was originally developed in 1965 to model citations connecting scientific papers [86]. This network is directed because a citation is a directed relationship: one paper cites another existing paper. If the average bibliography length of a paper is c citations, then c is the average out-degree of a node in the network. The model assumes that a new paper chooses old papers to cite with probability proportional to the number of citations the old paper already has. Since new papers must have zero citations (in-degree of zero), the model has to bend slightly to allow this to change. The probability of citing an existing paper is proportional to $k + r$, where k is the in-degree of the paper and r is a constant. This constant represents the rate that edges form uniformly at random. Typically when growing a network from scratch, we start with a network that is just two nodes connected by an edge.

The in-degree distribution for a network generated by the Price model follows a power-law distribution (high-degree nodes) [97]

$$f(k) \propto k^{-\alpha}$$

for $k \gg r$, where $\alpha = 2 + r/c$. This method is often referred to as linear preferential attachment because the probability of edge formation, or attachment, is proportional to the degree of a node.

Figure 2.5 shows three networks grown by linear preferential attachment. They all have $r = c = 1$, which means each node has out-degree of 1. As the network grows from 5 to 50 to 1000 nodes, we see the emergence of high-degree nodes. This property is one aspect of the structure that made this model so popular.

Note that if $2 < \alpha < 3$, the mean of the distribution of degrees for a node in the network is finite but the second moment is infinite, meaning the variance is infinite. Thus for a sufficiently large network, even though we have an expected value for the degree of a randomly chosen node, the infinite variance means we may see an enormous range of degree values. This is part of why these networks are often referred to as *scale free* [12]. The starting graph, that is the edges and nodes that we start with before beginning the algorithm, can influence the power-law behavior. The network has to grow sufficiently large to lose the impact of this original graph.

2.3.3 Vertex-copying model

Networks generated with the vertex-copying mechanism also produce scale-free structure. This method was inspired by the Price citation model [57], but proposes that new papers simply copy the entire bibliography of an existing paper. Thus the new vertex (new paper) essentially copies the existing vertex (old paper) by duplicating all of its edges.

In order to generate power-law degree distributions, the methods needs to allow new papers to gain citations and therefore needs a few modifications [78]. Instead of copying entire bibliographies, new papers will copy partial bibliographies and add the remaining citations uniformly at random.

The vertex-copying model assumes, like the Price model, that new nodes all have c out-going edges. When we add a new node, we choose uniformly at random which existing node to copy. We then go through each edge one at a time. For each edge, we either copy the edge to the new node with probability q , or with probability $1 - q$ we add an edge to another node in the network chosen uniformly at random.

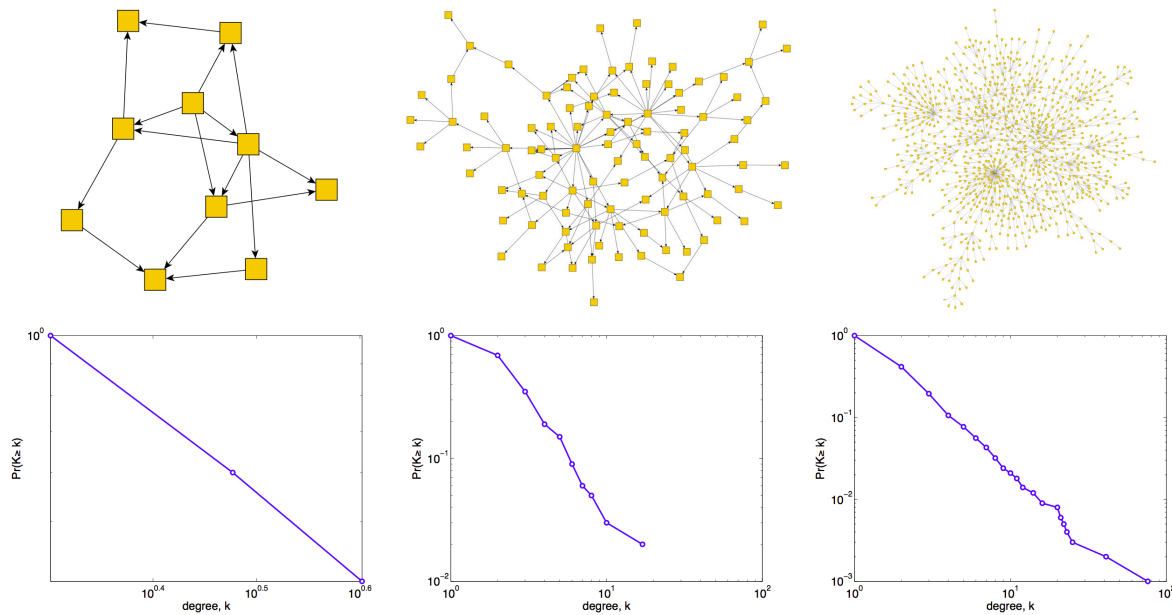


Figure 2.6: Networks grown by the vertex-copying mechanism with $q = 1/2$ and $n = 10, 100$, and 1000 . The blue lines are the empirical degree distributions. Reproduced from A. Clauset (2017) *Lecture 12: Growing Networks* [Lecture notes]. University of Colorado Boulder, CSCI 5352

As with preferential attachment, we do need to choose a starting graph, but as the network grows large enough, the properties of this graph become irrelevant. Asymptotically the model will generate power-law degree distributions. The power-law distribution that describes the degree distribution of a vertex-copy network has $\alpha = 1 + \frac{1}{q}$. Recall the q is the probability of copying an edge. Thus a completely copied network has $\alpha = 2$. As we increase the probability of uniform attachment, this exponent increases.

Figure 2.6 shows a series of vertex-copy graphs along with their empirical degree distributions. The size of the network increases from 10 to 100 to 1000 nodes, and q is set to $1/2$. Just like in the preferential attachment graphs, we see high-degree nodes appearing in the larger graphs. This is made more obvious by looking at the empirical degree distributions.

2.3.4 Configuration model

A configuration model generates networks with a given degree sequence. For example, we may want a network with a specific set of degrees, drawn from say an exponential, log-normal, or a power-law distribution. A configuration model allows us to generate such a network. Sometimes we may want to study what a specific network looks like if we rewire the edges and this too can be analyzed with a configuration model.

In Section 4.2, we use a configuration model to generate networks with a specified power-law degree sequence. We take the power-law degree sequence, \vec{k} and a number of nodes, n as input, and then use the Havel-Hakimi algorithm to connect the nodes as specified by \vec{k} [45, 43]. We then use a degree-preserving edge-swapping algorithm that can sample uniformly at random from the set of simple graphs with the same degree sequence [37]. The use of this algorithm after the Havel-Hakimi algorithm gives a more reasonable final output graph; sometimes the Havel-Hakimi algorithm generates graphs with unusual structure.

The Havel-Hakimi algorithm works as follows. We first check to make sure that pair (n, \vec{k}) is *graphical*, meaning that there exists a graph with the desired size and degree sequence. If $\max(\vec{k}) \geq n$, there does not exist a simple graph of size n with this degree sequence. Similarly, if the number of nodes that would be required to have odd degree (the number of odd values in \vec{k}) is not even, the sequence is not graphical. If the sequence passes these tests, the algorithm can begin.

We initialize as follows. Let x be a $2 \times n$ vector. The first row indexes $1, \dots, n$ and the second row is \vec{k} . We will denote the rows as $x[0]$ and $x[1]$, respectively.

Repeat the following while not all entries in $x[1]$ are zero:

- (1) If any entries in $x[1] < 0$, stop. The algorithm has failed and \vec{k} is not graphical.
- (2) Sort $x[1]$ in non-increasing order, so that $x[1, 0]$ is the largest value in $x[1]$.
- (3) Let k be this entry, $k = x[1, 0]$ and let $v = x[0, 0]$ correspond to the vertex in the network with this desired degree.

- (4) Now let $x = x[:, 1 :]$, removing the first column in the array that we just extracted. We wish to add k edges to vertex v .
- (5) To do this, we add edges $(v, x[0, i])$ for $i \in 0, 1, \dots, k - 1$. That is, we add an edge from v to the first k vertices in $x[0]$.
- (6) Now each of those vertices has gained an edge, so we subtract 1 from their desired degree count by setting $x[1, : k] = x[1, : k] - 1$.

The algorithm adds all the edges for each node in decreasing order of degree until all edges have been added. Note that the algorithm avoids self-loops by removing an entry from x before adding edges. It avoids multi-edges the same way: later nodes cannot form a connection back to a node that has already linked to it because the higher-degree node was already removed. Thus the algorithm gives a simple graph with the desired degree sequence, which we can feed into the edge-swapping algorithm [37] to give a perhaps more realistic or typical-looking graph.

Chapter 3

Prevalence of power laws in networks¹

We now turn to developing an initial test of the ubiquity of scale-free networks, which will build off the concepts, definitions, and tools described in Chapters 1 and 2, and develop new tests in order to construct a severe and robust test of the concept. Across the literature, the term “scale-free network” may mean a precise or approximate statistical pattern in the degree distribution, an emergent behavior in the asymptotic limit, or a property of all networks assembled in part or in whole by a particular family of mechanisms. This imprecision has contributed to the controversy around the scale-free hypothesis.

Here, we focus narrowly on the traditional degree-based definition of a scale-free network, which has the advantage of being directly testable using empirical data. However, even within this scope, the definition is often modified, depending on the setting, by introducing auxiliary hypotheses [66]. For instance, the scale-free pattern may only hold for the largest degrees, implying a formulation like $\Pr(k) \propto k^{-\alpha}$ for $k \geq k_{\min} > 1$. That is, the power law only appears in the upper tail of the distribution, and the lower tail or “body” follows some non-power-law pattern. In other settings, finite-size effects may suppress the frequency of nodes with degrees close to the underlying system’s size, implying a formulation like $\Pr(k) \propto k^{-\alpha} e^{-\lambda k}$ where λ models the system size. That is, the degrees follow a power-law distribution with an exponential cutoff in its extreme upper tail. Other times, heterogeneity in the degree distribution is of primary interest, implying a restriction on the parameter $\alpha \in (2, 3)$, where, in the asymptotic limit, the distribution’s mean is

¹ This chapter is adapted from: **A. D. Broido** and A. Clauset. Scale-free networks are rare. *Nature Communications* **10**: 1017 (2019)

finite while its variance is infinite. For some applications, the power law may not be meant as a good model of the data itself, but rather is claimed to be a better model than some alternatives, e.g., an exponential or a log-normal distribution, or merely represents the mathematical extreme of a “heavy-tailed” distribution, i.e., one that decays more slowly than an exponential.

A consequence of these varied uses of the term scale-free network, and its various auxiliary hypotheses, is that different researchers can use the same term to refer to slightly different concepts, and this ambiguity complicates any effort to empirically evaluate the basic hypothesis. Here, we construct a severe test [66] of the ubiquity of scale-free networks by applying state-of-the-art statistical methods to a large and diverse corpus of real-world networks. To explicitly cover the variations in how scale-free networks have been defined in the literature, we formalize a set of quantitative criteria that represent differing strengths and types of evidence for scale-free structure in a particular network. This set of criteria unifies the common variations, and their combinations, and allows us to assess different types and degrees of evidence of scale-free degree distributions.

For each network data set in the corpus, we estimate the best-fitting power-law model, test its statistical plausibility, and compare it to alternative non-scale-free distributions. We analyze these results collectively and consider how the evidence for scale-free structure varies across domains.

3.1 Data analysis pipeline

A key component of our evaluation of the scale-free hypothesis is the use of a large and diverse corpus of real-world networks. This corpus is composed of 928 network data sets drawn from the *Index of Complex Networks* (ICON), a comprehensive online index of research-quality network data, spanning all fields of science [23]. The composition of the corpus is roughly half biological networks, a third social or technological networks, and a sixth information or transportation networks (Table 3.1). The 928 networks included span five orders of magnitude in size, are generally sparse with a mean degree of $\langle k \rangle \approx 3$ (Fig. 3.1), and possess a range of graph properties, e.g., simple, directed, weighted, multiplex, temporal, or bipartite. These networks also exhibit a wide variety of graph properties.

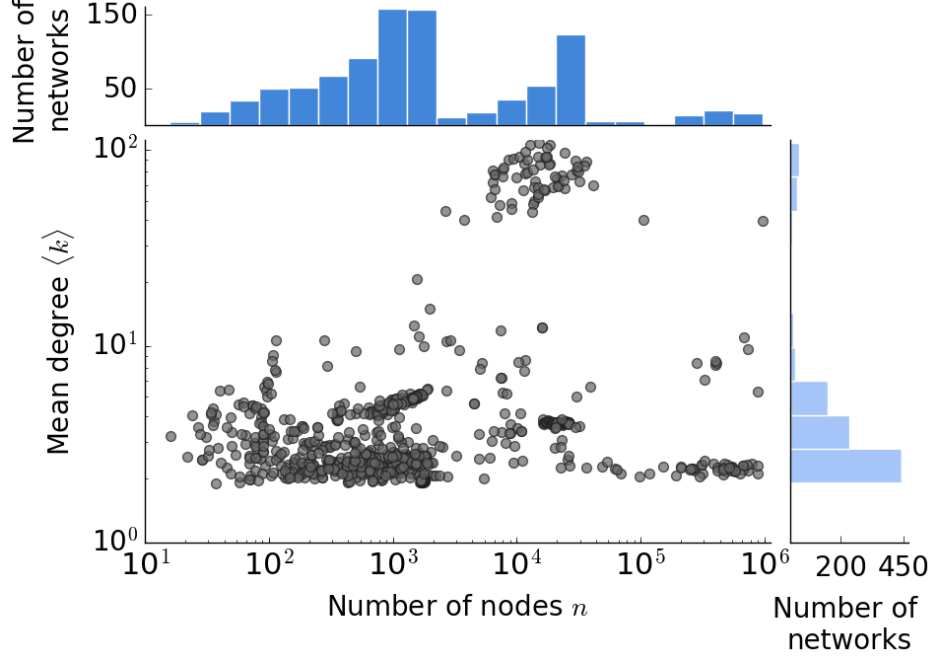


Figure 3.1: Mean degree $\langle k \rangle$ as a function of the number of nodes n . The 928 network data sets in the corpus studied here vary broadly size and density. For data sets with more than one degree sequence (see text), we plot the median of the corresponding set of mean degrees.

The scale-free hypothesis is defined most clearly for simple graphs, which have only one degree distribution. More complicated networks, e.g., a directed, weighted, multiplex network, can have multiple degree distributions, which complicates testing whether it is scale free: which degree distributions count as evidence and which do not? As our corpus includes both simple graphs and networks with various combinations of directed, weighted, bipartite, multigraph, temporal, and multiplex properties (Table 3.1), we must either solve this problem or restrict our analysis to only the simple networks.

To include these non-simple networks in our analysis we apply a sequence of graph transformations that convert a given *network data set*, defined as a network with multiple graph properties, into a set of simple graphs, each of which can be tested unambiguously for scale-free structure (Figs. 3.2). In this process, we discard any resulting simple graph that is either too dense or too sparse, under pre-specified thresholds, to be plausibly scale free (see 3.2 for details). Then, for each simple graph associated with a network data set, we apply standard statistical methods [22]

Domain	Number	(Prop.)	Multiplex	Bipartite	Multigraph	Weighted	Directed	Simple
Bio.	495	(0.53)	273	41	378	29	37	39
Info.	16	(0.02)	0	0	4	0	5	7
Social	147	(0.16)	7	0	6	8	0	129
Tech.	203	(0.22)	122	0	3	1	195	5
Trans.	67	(0.07)	48	0	65	3	2	0
Total	928	(1.00)	450	41	456	41	239	180

Table 3.1: Number of network data sets, and proportion of our network corpus, in each of five domains, under the taxonomy given by the *Index of Complex Networks* [23].

to identify the best-fitting power law in the degree distribution’s upper tail, evaluate its statistical plausibility using a goodness-of-fit test, and compare it to four alternative distributions fitted to the same part of the upper tail using a likelihood ratio test (see 3.3 and 3.4) . The outputs of these fitting, testing, and comparison procedures for a given simple graph encode in a vector the statistical evidence for its scale-free structure. We then evaluate the set of these vectors for a given network data set under criteria that formalize the different definitions of a scale-free network.

For a given degree distribution, a key step in this process is the selection of a value k_{\min} , above which the degrees are most closely modeled by a scale-free distribution (see 3.3). Hence, the fitting procedure truncates non-power-law behavior among low-degree nodes, enabling a more clear evaluation of potentially scale-free patterns in the upper tail. For technical reasons, all model tests and comparisons must then be made only on the degrees $k \geq k_{\min}$ in the upper tail [22]. Although our primary evaluation uses a normalized likelihood ratio test [107] that has been specifically shown valid for comparing the distributions considered here [22], we also present results based on using standard information criteria to compare distributional models [20] (see 4.1.6).

3.2 Extracting degree distributions from real-world networks

For each property, there can be multiple ways to extract a degree sequence, and in some cases, extracting a degree sequence requires making a choice. To resolve these ambiguities, we developed a set of graph simplification functions, which are applied in a sequence that depends only on the

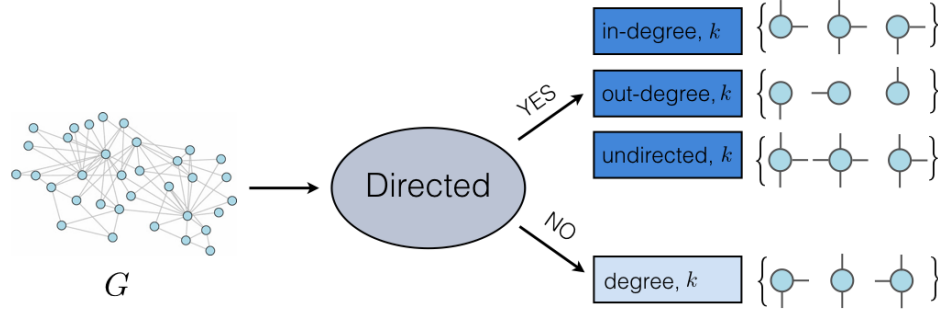


Figure 3.2: A graph simplification function, which takes as input a network G . In this case, if G is directed, the function returns three degree sequences: the in-degrees, out-degrees, and undirected degrees, while if G is undirected, it returns the degree sequence. 3.1 contains complete details.

graph properties of the input (Figures 3.2, 3.3). The purpose of this graph simplification algorithm is to provide an objective and consistent set of rules by which to extract a set of degree sequences from any given network data set. This approach thus removes researcher subjectivity in deciding which data set to include or exclude in any evaluation of the scale-free hypothesis, and ensures that the evaluation is as broad as possible. For completeness, we describe these specific pathways, and give counts of how many network data sets in our corpus followed each pathway.

At each stage in our processing we remove one graph property, making the network simpler and never adding properties. Repeating this process for each property in succession converts a network data set into a set of simple graphs. Some networks are processed into a large number of simple graphs, due to the combinatoric effect of certain graph properties. To moderate the amount of combinatoric blowup, we treat weighted graphs differently depending on whether or not they have any multiplex, bipartite, or multigraph properties. Multiplex networks include temporal networks as a special case; many of these have a large number of layers, each of which can generate many simple graphs (see below).

If a weighted graph has any of the aforementioned properties, we simply ignore the edge weights and process the remaining properties. If not, however, the data set is replaced with three unweighted graphs as follows. The goal of this transformation is to replace a potentially dense weighted graph, e.g., a data set representing pairwise similarity scores or correlations, with a set

of unweighted graphs that are relatively sparse. To carry out this conversion, we choose thresholds intended to produce sparse graphs that are not so sparse as to be too strongly disconnected to be potentially scale free. Toward this end, we identify and then apply three thresholds to the edge weights, so that the resulting unweighted graphs have a mean degree $\langle k \rangle = \{2, n^{1/4}, \sqrt{n}\}$. These threshold values are determined by the empirical edge weight distribution of the graph, and correspond to choosing the $m = \{n, (1/2)n^{5/4}, (1/2)n^{3/2}\}$ largest-weight edges, respectively. The lower value of $\langle k \rangle$ or m produces a very sparse graph, retaining primarily the largest-weight edges, but not so sparse as to be likely strongly disconnected. The upper value produces a more well connected network, retaining all but the smallest-weight edges, but not so dense that the degree distribution is trivial. The middle value splits the difference between these. Our corpus contains only 8 weighted networks and 6 weighted directed networks for 14 total weighted networks, meaning that these networks represent a modest share (2%) of the corpus.

Multiplex and temporal network data sets are composed of T “layers,” each of which is a network itself. The multiplex network is replaced by a set of $T + 1$ graphs, one for each layer and one for the union of edges and nodes across all layers. In this way, the multiplex or temporal property is removed, and the original data set replaced with a set of graphs. Each graph in this set is then further processed to remove any remaining non-simple properties. A bipartite graph is replaced with three graphs: one each for the “A-mode” projection, “B-mode” projection, and original bipartite graph. If present, multi-edges are collapsed and weights discarded.

As a final step, directed graphs are replaced by three degree sequences: one for the in-degrees, one for the out-degrees, and one for the total degrees; undirected graphs are replaced with their single degree sequence. The results of this sequential processing is a set of degree sequences that, as a group, represent the original network. Our corpus contains 5 pure multiplex networks, 315 multiplex multigraphs, and 130 multiplex directed networks, which yields 450 total multiplex networks.

Network data sets that are bipartite and not multiplex are first replaced with three graphs: one for the “A-mode” project, one for the “B-mode” projection, and one for the original bipartite

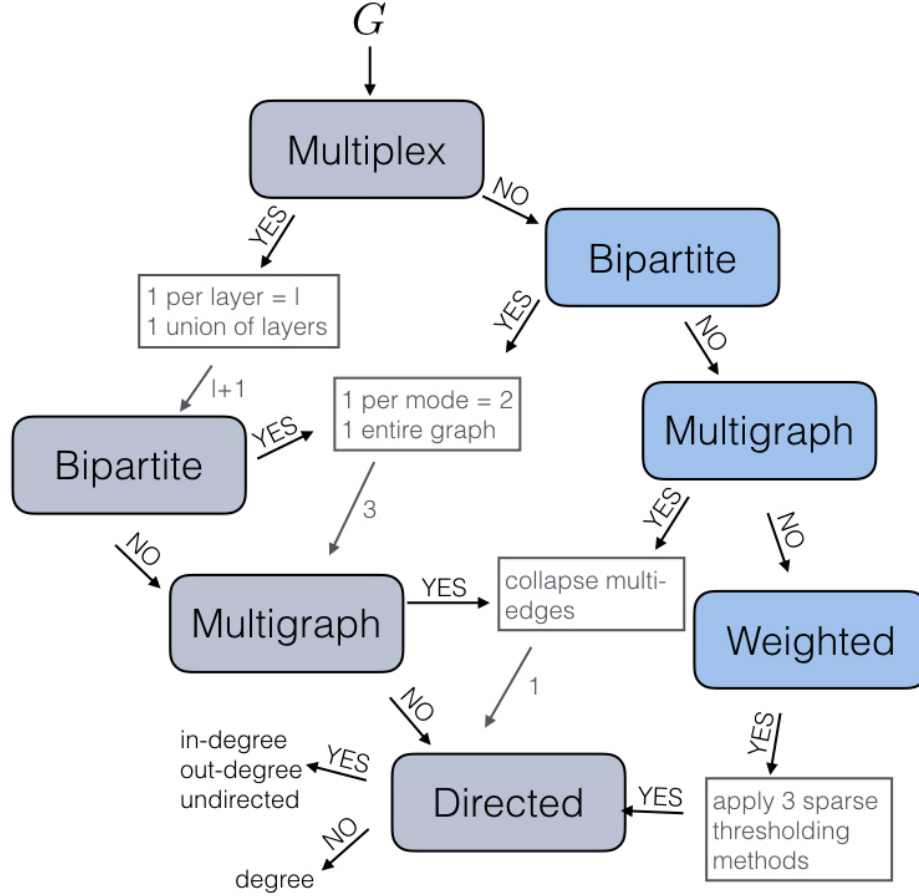


Figure 3.3: Flowchart describing the path from network data set to degree sequence(s). Each step removes a layer from the properties. The gray path is for multiplex, bipartite, or multigraph networks, while the blue is for weighted networks without these properties. Details in text.

graph. Each of these graphs is then processed starting from just after the bipartite step described above in the multiplex or temporal network processing pathway. In our corpus, there are 16 purely bipartite networks, and 25 bipartite weighted networks, which yields 41 bipartite networks total (4% of the corpus).

Data sets that are multigraphs, but not multiplex/temporal or bipartite, are merely simplified by collapsing multi-edges. Edge weights are then discarded, and the resulting graph is processed starting from the check for directedness as above. In our corpus, there are 139 multigraphs and 2 weighted multigraphs, which yields 456 multigraphs total, including those that are multiplex.

Data sets that are only directed, with no other properties, are processed to produce three

degree sequences: one each for the in-degrees, out-degrees, and total degrees. In our corpus, there are 103 purely directed networks (11.1% of data sets). In the case of a simple graph, the degree sequence is taken with no further processing. Our corpus contains 180 simple networks (19.4% of data sets).

Complicated network data sets can produce a combinatoric number of simple graphs under this process. Treating every simplified degree sequence independently could lead to skewed results, e.g., if a few non-scale-free data sets account for a large fraction of the total extracted simple graphs. To avoid this bias, results are reported at the level of network data sets. Additionally, we require that simplified graphs are neither too sparse nor too dense to be potentially scale free and thus retain for analysis only simplified graphs with mean degree $2 < \langle k \rangle < \sqrt{n}$.

Simplifying the 928 network data sets produced 18,448 simple graphs, of which 14,415 were excluded for being too sparse and 371 excluded for being too dense (about 80.4% of derived simple graphs). Results are reported only in terms of the remaining 3662 simple graphs (about 3.9 per network data set). Of the 928 network data sets, 735 (79%) produced no graphs that were excluded for being too sparse. More than 90% of graphs excluded for being too sparse were produced by simplifying 3 network data sets (less than 1% of the corpus). Similarly, 874 (94%) of the network data sets produced no graphs that were excluded for being too dense. More than 70% of graphs excluded for being too dense were produced by simplifying 3 network data sets. Finally, 782 (84%) of the data sets generated at most 3 degree sequences prior to applying the too-sparse and too-dense filters. Hence, the vast majority of data sets were uninvolved in the production of many excluded graphs.

This approach for evaluating evidence for scale-free structure has several advantages. It provides a systematic procedure applicable to any network data set, and treats every data set equivalently. It provides an evaluation of the scale-free hypothesis over a maximally broad variety of networks, which facilitates the characterization of their empirical ubiquity. And, it provides a means to assess different kinds of evidence for scale-free structure, by combining results from multiple degree distributions, if available in a network data set. The graph-simplification process

or the particular evidence criteria used may also introduce biases into the results. We control for these possibilities by considering alternative criteria under multiple robustness analyses.

3.3 Power-law analysis

For the degree sequence $\{k_i\} = k_1, k_2, \dots, k_n$ of a given network data set, we estimate the best-fitting power-law distribution of the form

$$\Pr(k) = C k^{-\alpha} \quad \alpha > 1, \quad k \geq k_{\min} \geq 1, \quad (3.1)$$

where α is the scaling exponent, C is the normalization constant, and k is integer valued. This specification models only the distribution's upper tail, i.e., degree values $k \geq k_{\min}$, and discards data from any non-power-law portion in the lower distribution.

3.3.1 Modeling power-law degree distributions

As the data in a degree sequence is integer-valued, starting at $k_{\min} \geq 1$, then pmf of the power law has the form

$$\Pr(k) = \frac{1}{\zeta(\alpha, k_{\min})} k^{-\alpha}$$

where $\zeta(\alpha, k_{\min}) = \sum_{i=0}^{\infty} (i + k_{\min})^{-\alpha}$ is the Hurwitz zeta function.

Fitting this model to an empirical degree sequence requires first choosing the location \hat{k}_{\min} at which the upper tail begins, and then estimating the scaling exponent $\hat{\alpha}$ on the truncated data $k \geq \hat{k}_{\min}$. Because the choice of k_{\min} changes the sample size, it cannot be directly estimated using likelihood or Bayesian techniques. Here, the standard KS-minimization approach is used to choose \hat{k}_{\min} and the discrete maximum likelihood estimator is used to choose $\hat{\alpha}$ [22]. This method selects the k_{\min} that minimizes the maximum difference in absolute value between the (cumulative) empirical distribution $E(k)$ on the observed degrees $k \geq k_{\min}$ and the cmf of the best fitting power law $P(k | \hat{\alpha})$ on those same observations. This difference, called the KS statistic, is defined as

$$D = \max_{k \geq k_{\min}} |E(k) - P(k | \hat{\alpha})|.$$

We choose as k_{\min} the value that minimizes D . The estimate $\hat{\alpha}$ is chosen by maximum likelihood (the MLE), which we obtain by numerically optimizing the log-likelihood function [22].

3.3.2 Testing goodness-of-fit

Fitting the power-law distribution always returns some parameters $\hat{\theta} = (\hat{k}_{\min}, \hat{\alpha})$. However, parameters alone give no indication of the quality of the fitted model. We assess the goodness-of-fit of the fitted model using a standard p -value, numerically estimated via a semi-parametric bootstrap approach [22]. Given a degree sequence with n elements, of which n_{tail} are $k \geq k_{\min}$ and with MLE $\hat{\alpha}$, a synthetic data set is generated as follows. For each of n synthetic values, with probability n_{tail}/n we draw a random deviate from the fitted power-law model, with parameters k_{\min} and $\hat{\alpha}$. Otherwise, we choose a value uniformly at random from the empirical set of degrees $k < k_{\min}$. Repeated n times this produces a synthetic data set that closely follows the empirical distribution below k_{\min} and follows the fitted power-law model at and above k_{\min} .

Applying the previously defined fitting procedure to a large number of these synthetic data sets yields the null distribution of the KS-statistic $\Pr(D)$. Let D^* denote the value of the KS-statistic for the best fitting power-law model for the empirical degree sequence. The p -value for this model is defined as the probability of observing, under the null distribution, a KS-statistic at least as extreme as D^* . Hence, $p = \Pr(D \geq D^*)$ is the fraction of synthetic data sets with KS statistic larger than that of the empirical data set. Following standard practice for power-law degree distributions [22], if $p < 0.1$, then we reject the power law as a plausible model of the degree sequence, and if $p \geq 0.1$, then we fail to reject the model. We note: failing to reject does not imply that the model is correct, only that it is a plausible data generating process. Hence, if the underlying data generating process is indeed scale free, this test has a false negative rate of 0.1. The results of this test provide direct evidence for or against a network exhibiting scale-free structure.

3.4 Alternative Distributions

Each power-law model $\hat{\theta}$ is compared to four non-scale-free alternative models. These are estimated via maximum likelihood on the same degrees $k \geq \hat{k}_{\min}$ as found in the best power-law fit so as to be comparable with a likelihood ratio test (see 3.4.5). The non-scale free alternatives used here are the (i) exponential, (ii) log-normal, (iii) power-law with exponential cutoff, and (iv) stretched exponential or Weibull distributions, all of which have been used previously as models of degree distributions [9, 18, 55, 64, 33], and for which the validity of the LRT used here has specifically been previously established [22].

3.4.1 Exponential

The exponential distribution looks like $\Pr(k) \propto e^{-\lambda k}$. To start at k_{\min} , the normalization constant must be

$$\begin{aligned} \sum_{k=k_{\min}}^{\infty} \Pr(x) &= \sum_{k=k_{\min}}^{\infty} C e^{-\lambda k} = 1 \\ \implies 1/C &= \sum_{k=k_{\min}}^{\infty} e^{-\lambda k} = (1 - e^{-\lambda}) e^{\lambda k_{\min}} \end{aligned}$$

Thus the pmf of the discrete exponential has the form

$$\Pr(k) = \left(\frac{e^{-\lambda k_{\min}}}{1 - e^{-\lambda}} \right) e^{-\lambda k} .$$

As with the power-law distribution, we use standard numerical maximization routines to estimate the maximum likelihood choice of λ .

3.4.2 Log-normal

The log-normal distribution is typically defined on a continuous variable k . To adapt this distribution to discrete values, we bin the continuous distribution and then adjust so that it begins at k_{\min} rather than at 0.

Let $f(k)$ and $F(k)$ be the density and distribution functions of a continuous log-normal variable, where

$$f(k) = \frac{1}{\sqrt{2\pi}\sigma k} e^{-\frac{(\log k - \mu)^2}{2\sigma^2}} \quad , \quad x > 0$$

and

$$F(k) = \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left[\frac{(\log k - \mu)}{\sqrt{2}\sigma} \right] \quad .$$

We define $g(k)$ and $G(k)$ to be the density and distribution functions of a discrete log-normal variable, given by

$$g(k) = F(k+1) - F(k) \quad , \quad x \geq 0$$

and

$$G(k) = \sum_{y=0}^k g(y) = F(k+1) - F(0) = F(k+1) \quad .$$

We then generalize the distribution to start at some minimum value, i.e., rather than starting at 0, the distribution starts at $k = k_{\min}$, where k_{\min} is a positive integer. This pmf is obtained by re-normalizing the tail of $g(k)$ so that it sums to 1 on the interval k_{\min} to ∞ , yielding

$$\begin{aligned} h(k) &= \frac{g(k)}{\sum_{k=k_{\min}}^{\infty} g(k)} = \frac{g(k)}{1 - \sum_{k=0}^{k_{\min}-1} g(k)} \\ &= \frac{g(k)}{1 - G(k_{\min} - 1)} = \frac{g(k)}{1 - F(k_{\min})} \quad . \end{aligned}$$

Maximum likelihood estimation was carried out using standard numerical optimization routines. Additionally, we constrained the optimization in order to prevent numerical instabilities. Specifically, we required $\sigma \geq 1$ and $\mu \geq -\lfloor n/5 \rfloor$. As a check on these constraints, we verified that in no cases did the likelihood improve significantly by allowing $\sigma < 1$, and the constraint on μ prevents it from decreasing without bound (a behavior that can produce arbitrarily heavy-tailed distributions over a finite range in the upper tail). To initialize the numerical search, we set $(\mu_0, \sigma_0) = (0, 1)$. As the data we fit is strictly from the tail, it seems likely that the mean in particular would be an overestimate of the mean of our target distribution. Since μ converges to negative values, 0 is a conservative initial guess.

3.4.3 Power-law with exponential cutoff

The power-law with exponential cutoff is a combination of an exponential and power-law distribution, with $\Pr(k) \propto k^{-\alpha} e^{-\lambda k}$. The normalization constant for a distribution beginning at k_{\min} is found as follows:

$$\begin{aligned} \sum_{k=k_{\min}}^{\infty} \Pr(k) &= \sum_{k=k_{\min}}^{\infty} C k^{-\alpha} e^{-\lambda k} = 1 \\ \implies 1/C &= \sum_{k=k_{\min}}^{\infty} k^{-\alpha} e^{-\lambda k} \\ &= e^{-k_{\min} \lambda} * \Phi(e^{-\lambda}, \alpha, k_{\min}) \end{aligned}$$

where $\Phi(z, s, a) = \sum_{i=0}^{\infty} \frac{z^i}{(a+i)^s}$ is the Lerch Phi function. The pmf then has the form

$$\Pr(k) = \left[e^{-k_{\min} \lambda} \Phi(e^{-\lambda}, \alpha, k_{\min}) \right]^{-1} k^{-\alpha} e^{-\lambda k}$$

We estimate this distribution's parameters λ and α using standard numerical maximization routines.

3.4.4 Weibull (Stretched exponential)

A common approach to obtain a discrete version of the stretched exponential or Weibull distribution is to bin the continuous distribution [72]. Let $f(k)$ and $F(k)$ be the density and distribution functions of a continuous Weibull variable, where

$$F(k) = 1 - e^{-(k/b)^a}, \quad x \geq 0 .$$

Define $g(k)$ and $G(k)$ to be the density and distribution functions of a discrete Weibull variable, given by:

$$g(k) = F(k+1) - F(k), \quad x \geq 0$$

and

$$G(k) = \sum_{y=0}^k g(y) = F(k+1) - F(0) = F(k+1) .$$

As with the log-normal, we generalize the distribution to start at some minimum value, i.e., rather than starting at 0, the distribution starts at $k = k_{\min}$, where k_{\min} is a positive integer. This pmf is obtained by re-normalizing the tail of $g(k)$ so that it sums to 1 on the interval k_{\min} to ∞ , yielding

$$\begin{aligned} h(k) &= \left[\sum_{k=k_{\min}}^{\infty} g(k) \right]^{-1} g(k) = \left[1 - \sum_{k=0}^{k_{\min}-1} g(k) \right]^{-1} g(k) \\ &= [1 - G(k_{\min} - 1)]^{-1} g(k) = [1 - F(k_{\min})]^{-1} g(k) \\ &= e^{(k_{\min}/b)^a} \left[e^{-(k/b)^a} - e^{-((k+1)/b)^a} \right] . \end{aligned}$$

We estimate this distribution's parameters using standard numerical maximization routines.

3.4.5 Likelihood-ratio tests

Each power-law model $\hat{\theta}$ is compared to the four non-scale-free alternative models, estimated via maximum likelihood on the same degrees $k \geq \hat{k}_{\min}$, using a standard Vuong normalized likelihood ratio test (LRT) [22, 107]. These likelihood ratio tests have been previously shown valid for both the nested and non-nested models considered here [107, 22], and have lower incorrect decision rates [22] compared to simple penalized likelihood approaches to model comparison. The restriction to $k \geq \hat{k}_{\min}$ is necessary to make the model likelihoods directly comparable, and slightly biases the test in favor of the power law, as the best choice of \hat{k}_{\min} for an alternative may not be the same as the best choice for the power law [22]. The results of this test provide indirect evidence about the scale-free hypothesis, as a power-law model can be favored over some alternative even if the power law itself is not a statistically plausible model of the data.

For each alternative distribution, we obtain the log-likelihood \mathcal{L}_{Alt} of the best fit. The difference between this value and the log-likelihood of the power-law fit to the same observations yields the likelihood ratio test (LRT) statistic

$$\mathcal{R} = \mathcal{L}_{\text{PL}} - \mathcal{L}_{\text{Alt}}$$

where \mathcal{L}_{PL} is the log-likelihood of the power-law model. We calculate this statistic for each alternative model.

The sign of \mathcal{R} indicates which model is a better fit to the data. When $\mathcal{R} > 0$, the power law is a better fit to the data, and when $\mathcal{R} < 0$, the alternative distribution is the better fitting model. Crucially, when $\mathcal{R} = 0$, the test is inconclusive, meaning that the data cannot distinguish between the two models.

The test statistic \mathcal{R} is derived from data, meaning that it is itself a random variable subject to statistical fluctuations [22, 107]. Accounting for these fluctuations dramatically improves the accuracy of the test by reducing both types of incorrect decision rates [22]. As a result, the sign of \mathcal{R} alone is not a reliable indicator of which model is a better fit. Specifically, the sign of \mathcal{R} is meaningful only if its magnitude $|\mathcal{R}|$ is statistically distinguishable from 0. This determination is made by a standard two-tailed test against a null hypothesis of $\mathcal{R} = 0$, which yields a standard p -value. If $p \geq 0.1$, then $|\mathcal{R}|$ is statistically indistinguishable from 0 and neither model is a better explanation of the data than the other. If $p < 0.1$, then the data provide a clear conclusion in favor of one model or the other, depending on the sign of \mathcal{R} . This threshold sets the false positive rate for the alternative distribution at 0.05. Corrections for multiple tests, e.g., a family-wise error rate method like Bonferroni or a false discovery correction like Benjamini-Hochberg, are not employed. Such corrections would simply lower the obtained p -values without changing the overall conclusions, while introducing additional assumptions into the analysis.

We obtain this p -value with the same method used in Ref. [22], originally proved valid in Ref. [107]. Note that

$$\begin{aligned}\mathcal{R} &= \mathcal{L}_{\text{PL}} - \mathcal{L}_{\text{Alt}} \\ &= \sum_{i=1}^n [\ln \text{Pr}_{\text{PL}}(k_i) - \ln \text{Pr}_{\text{Alt}}(k_i)] \\ &= \sum_{i=1}^n [\ell_i^{(\text{PL})} - \ell_i^{(\text{Alt})}]\end{aligned}$$

where $\ell_i^{(\text{PL})}$ is the log-likelihood of a single observed degree value k_i under the power-law model,

and n is the number of empirical observations being used by a model (in our setting, this number is n_{tail} , but we omit that annotation to keep the mathematics more compact).

We have assumed that the degree values k_i are independent, which means the point-wise log-likelihood ratios $\ell_i^{(\text{PL})} - \ell_i^{(\text{Alt})}$ are independent as well. The central limit theorem states that the normalized sum of independent random variables becomes approximately normally distributed as their number grows large, and that this normal distribution has mean μ and variance $n\sigma^2$, where σ^2 is the variance of a single term. This distribution can be used to obtain the p -value, but requires that we first estimate μ and σ^2 . Note that we assume $\mu = 0$ because the null hypothesis is $\mathcal{R} = 0$. We then approximate σ^2 as the sample variance in the observed \mathcal{R}

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n \left[\left(\ell_i^{(\text{PL})} - \ell_i^{(\text{Alt})} \right) - \left(\bar{\ell}^{(\text{PL})} - \bar{\ell}^{(\text{Alt})} \right) \right]^2 ,$$

where

$$\bar{\ell}^{(\text{PL})} = \frac{1}{n} \sum_{i=1}^n \ell_i^{(\text{PL})} \quad \text{and} \quad \bar{\ell}^{(\text{Alt})} = \frac{1}{n} \sum_{i=1}^n \ell_i^{(\text{Alt})}$$

are sample means.

Under this null distribution, the probability of observing an absolute value of \mathcal{R} at least as large as the actual test statistic is given by the two-tail probability

$$p = \frac{1}{\sqrt{2\pi n\sigma^2}} \left[\int_{-\infty}^{-|\mathcal{R}|} e^{-\frac{t^2}{2n\sigma^2}} dt + \int_{|\mathcal{R}|}^{\infty} e^{-\frac{t^2}{2n\sigma^2}} dt \right]. \quad (3.2)$$

Hence, following standard practice [22], if $p \leq 0.1$, then we reject the null hypothesis that $\mathcal{R} = 0$, and proceed by interpreting the sign of \mathcal{R} as evidence in favor of one or the other model.

To report results at the level of a network data set, we apply the LRTs to all the associated simple graphs and then aggregate the results. For each alternative distribution, we count the number of simple graphs associated with a particular network data set in which the outcome favored the alternative, favored the power law, or had an inconclusive result. Normalizing these counts across scale-free categories provides a continuous measure of the relative evidence that the data set falls into each of category.

3.5 Definitions of a scale-free network.

The different notions of evidence for scale-free structure found in the literature can be organized into a nearly nested set of categories (Fig. 3.4) and assessed by applying standard statistical tools to each graph associated with a network data set. Evidence for scale-free structure typically comes in two types: (i) a power-law distribution is not necessarily a good model of the degrees, but it is a relatively better model than alternatives, or (ii) a power law is itself a good model of the degrees.

The first type represents indirect evidence of scale-free structure, because the observed degree distribution is not itself required to be plausibly scale free, only that a scale-free pattern is more believable than some non-scale-free patterns. A network data set that exhibits this kind of evidence is placed into a category called

Super-Weak: For at least 50% of graphs, no alternative distribution is favored over the power law.

The second type represents direct evidence of scale-free structure, and the various modifications of a purely scale-free pattern can be organized in a set of nested categories that represent increasing levels of evidence:

Weakest: For at least 50% of graphs, a power-law distribution cannot be rejected ($p \geq 0.1$).

Weak: Requirements of *Weakest*, and the power-law region contains at least 50 nodes ($n_{\text{tail}} \geq 50$).

Strong: Requirements of *Weak* and *Super-Weak*, and $2 < \hat{\alpha} < 3$ for at least 50% of graphs.

Strongest: Requirements of *Strong* for at least 90% of graphs, and requirements of *Super-Weak* for at least 95% of graphs.

The progression from Weakest to Strongest categories represents the addition of more specific properties of the power-law degree distribution, all found in the literature on scale-free networks

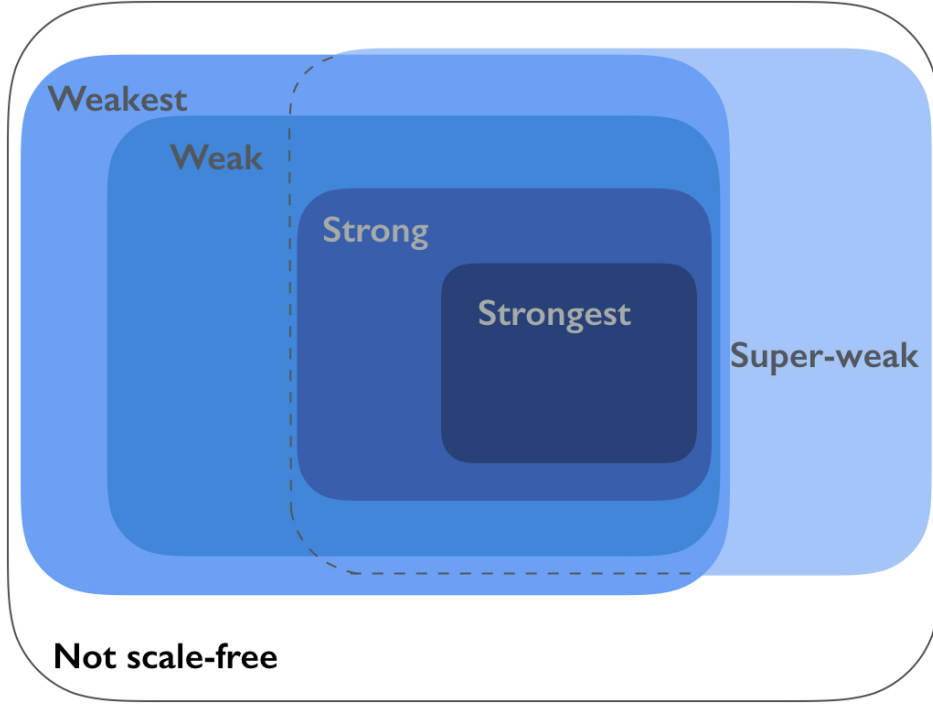


Figure 3.4: Taxonomy of scale-free network definitions. *Super-Weak* meaning that a power law is not necessarily a statistically plausible model of a network’s degree distribution but it is less implausible than alternatives; *Weakest*, meaning a degree distribution that is plausibly power-law distributed; *Weak*, adds a requirement that the distribution’s scale-free portion cover at least 50 nodes; *Strong*, adds a requirement that $2 < \hat{\alpha} < 3$ and the *Super-Weak* constraints; and, *Strongest*, meaning that every associated simple graph can meet the *Strong* constraints. The *Super-Weak* overlaps with the *Weak* definitions and contains the *Strong* definitions as special cases. Networks that fail to meet any of these criteria are deemed Not Scale Free.

or distributions. We define a sixth category of networks that includes all networks that do not fall into any of the above categories:

Not Scale-Free: Networks that are neither Super-Weak nor Weakest.

This evaluation scheme is parameterized by the different fractions of simple graphs required by each evidence category. The particular thresholds given above are statistically motivated in order to control for false positives and overfitting, and to provide a consistent treatment across all networks. The above scheme favors finding evidence for scale-free structure in three ways: (i) graphs identified as being too dense or too sparse to be plausibly scale free are excluded from all analyses, (ii) the estimation procedure selects, by choosing k_{\min} , the subset of data in the upper tail that

best-fits a power law, and (iii) the comparisons to alternatives are performed only on the data selected by the power law.

3.5.1 Parameters for defining scale-free networks

Threshold parameters for the primary evaluation criteria were selected to balance false positive and false negative rates, and to provide a consistent evaluation of evidence independent of the associated graph properties or source of data. For the Super-Weak and Weakest categories, a threshold of 50% ensures that the given property is present in a majority of simple graphs associated with a network data set. For the Weak category, a threshold of at least 50 nodes covered by the best-fitting power law in the upper tail follows standard practices [22] to reduce the likelihood of false positive errors due to low statistical power. For the Strong category, $\alpha \in (2, 3)$ covers the full parameter range for which scale-free distributions have an infinite second moment but a finite first moment. For the Strongest category, the thresholds of 90% for the goodness-of-fit test and 95% for likelihood ratio tests against alternatives match the expected error rates for both tests under the null hypothesis. If every graph associated with a network data set is scale free, the goodness-of-fit test is expected to incorrectly reject the power-law model 0.1 of the time, and the likelihood ratio test will falsely favor the alternative 0.05 of the time.

For specific networks, domain knowledge may suggest that some degree sequences are potentially scale free while others are likely not. A non-uniform weighting scheme on the set of associated degree sequences would allow such prior knowledge to be incorporated in a Bayesian fashion. However, no fixed non-uniform scheme can apply universally correctly to networks as different as, for example, directed trade networks, directed social networks, and directed biological networks. To provide a consistent treatment across all networks, regardless of their properties or source, we employ an uninformative (uniform) prior, which assigns equal weight to each associated degree sequence. In future work on specific subgroups of networks, a domain-specific weight scheme could be used with the evaluation criteria described here.

3.6 Empirical results

3.6.1 Scaling parameters.

Across the corpus, the distribution of median estimated scaling parameters $\hat{\alpha}$ is concentrated around a value of $\hat{\alpha} = 2$, but with a long right-tail such that 32% of data sets exhibit $\hat{\alpha} \geq 3$ (Fig. 3.5). The range $\alpha \in (2, 3)$ is sometimes identified as including the most emblematic of scale-free networks [11, 30], and we find that 39% of network data sets have median estimated parameters in this range. We also find that 34% of network data sets exhibit a median parameter $\hat{\alpha} < 2$, which is a relatively unusual value in the scale-free network literature.

Because every network produces some $\hat{\alpha}$, regardless of the statistical plausibility of the network being scale free, the shape of the distribution of $\hat{\alpha}$ is not necessarily evidence for or against the ubiquity of scale-free networks. It does, however, enable a check of whether the estimation methods are biased by network size n . Comparing $\hat{\alpha}$ and n , we find little evidence of strong systematic bias ($r^2 = 0.24$, $p = 1.82 \times 10^{-13}$; Fig. 3.6).

Across the five categories of evidence for scale-free structure, the distribution of median $\hat{\alpha}$ parameters varies considerably (Fig. 3.5, insets). For networks that fall into the Super-Weak category, the distribution has a similar breadth as the overall distribution, with a long right-tail and many networks with $\hat{\alpha} > 3$. Most of the networks with $\hat{\alpha} < 2$ are spatial networks, representing mycelial fungal or slime mold growth patterns [59]. However, few of these exhibit even Super-Weak or Weakest evidence of scale-free structure, indicating that they are not particularly plausible scale-free networks. Among the Weakest and Weak categories, the distribution of median $\hat{\alpha}$ remains broad, with a substantial fraction exhibiting $\hat{\alpha} > 3$. The Strong and Strongest categories require that $\hat{\alpha} \in (2, 3)$, and the few network data sets in these categories are somewhat concentrated near $\hat{\alpha} = 2$.

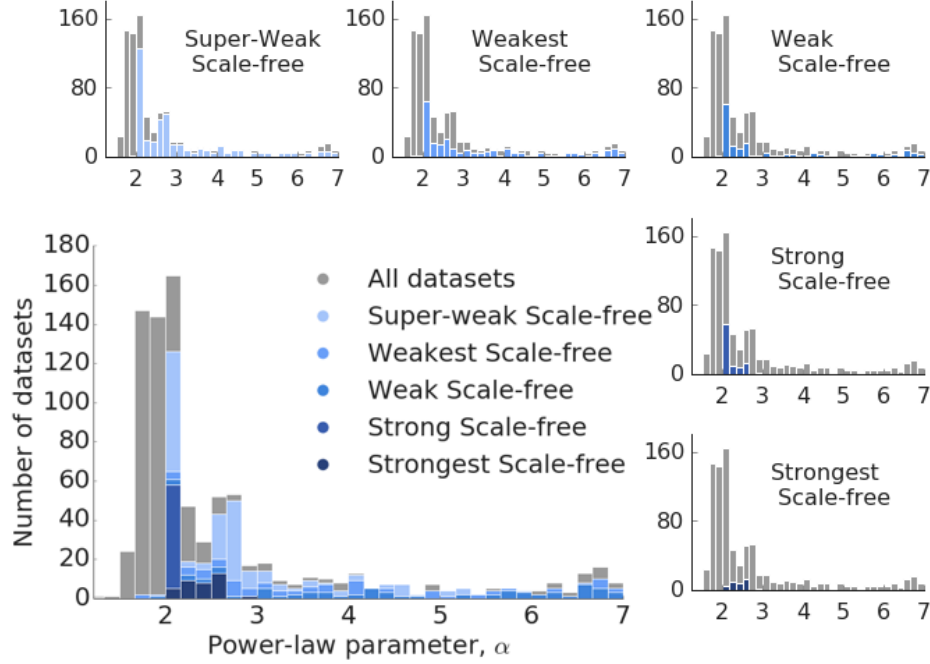


Figure 3.5: Distribution of $\hat{\alpha}$ by scale-free evidence category. For networks with more than one degree sequence, the median estimate is used, and for visual clarity the 8% of networks with a median $\hat{\alpha} \geq 6.5$ are omitted.

3.6.2 Alternative distributions.

Independent of whether the power-law model is a statistically good model of a network's degree sequence, it may nevertheless be a better model than non-power-law alternatives.

Across the corpus, likelihood ratio tests find only modest support for the power-law distribution over four alternatives (Table 3.2). In fact, the exponential distribution, which exhibits a thin tail and relatively low variance, is favored over the power law (41%) more often than vice-versa (33%). This outcome accords with the broad distribution of scaling parameters, as when $\alpha > 3$ (32% of data sets; Fig. 3.5), the degree distribution must have a relatively thin tail.

The log-normal is a broad distribution that can exhibit heavy tails, but which is nevertheless not scale free. Empirically, the log-normal is favored more than three times as often (48%) over the power law, as vice versa (12%), and the comparison is inconclusive in a large number of cases (40%). In other words, the log-normal is at least as good a fit as the power law for the vast majority

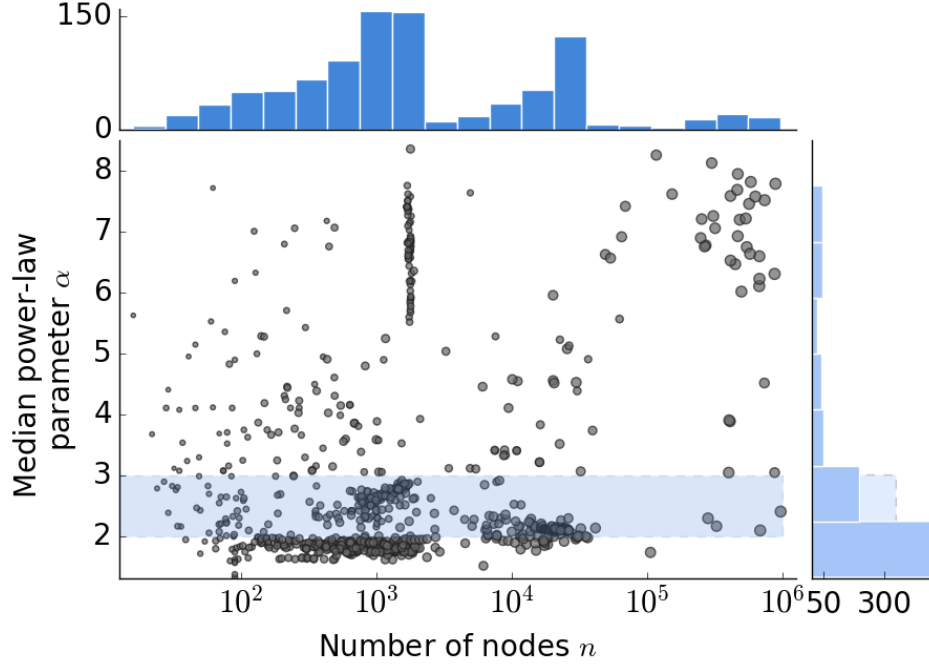


Figure 3.6: Median $\hat{\alpha}$ parameter versus network size n . A horizontal band highlights the canonical $\alpha \in (2, 3)$ range and illustrates the broad diversity of estimated power-law parameters across empirical networks.

of degree distributions (88%), suggesting that many previously identified scale-free networks may in fact be log-normal networks.

The Weibull or stretched exponential distribution can produce thin or heavy tails, and is a generalization of the exponential distribution. Compared to the power law, the Weibull is more often the better statistical model (47%) than vice versa (33%). Finally, the power-law distribution with an exponential cut-off requires special consideration, as it contains the pure power-law model as a special case. As a result, the likelihood of the power law can never exceed that of the cutoff model, and the interesting outcome is the degree to which the test is inconclusive between the two. In this case, a majority of networks (56%) favor the power law with cutoff model, indicating that finite-size effects may be common.

The above findings are corroborated by replacing the likelihood ratio test with information criteria to perform the model comparisons, which yield qualitatively similar conclusions (Table 4.1).

Alternative	$p(x) \propto f(x)$	Test Outcome		
		M_{PL}	Inconclusive	M_{Alt}
Exponential	$e^{-\lambda x}$	33%	26%	41%
Log-normal	$\frac{1}{x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$	12%	40%	48%
Weibull	$e^{-\left(\frac{x}{b}\right)^a}$	33%	20%	47%
Power law with cutoff	$x^{-\alpha} e^{-\lambda x}$	—	44%	56%

Table 3.2: Comparison of scale-free and alternative distributions. The percentage of network data sets that favor the power-law model M_{PL} , alternative model M_{Alt} , or neither, under a likelihood-ratio test, along with the form of the alternative distribution $f(x)$.

3.6.3 Assessing the scale-free hypothesis.

Given the results of fitting, testing, and comparing the power-law distribution across networks, we now classify each according to the six categories described above.

Across the corpus, fully 49% of networks fall into the Not Scale Free category (Fig. 3.7). Slightly less than half (46%) fall into the Super-Weak category, in which a scale-free pattern among the degrees is not necessarily statistically plausible itself, but remains no less plausible than alternative distributions. The Weakest and Weak categories represent networks in which the power-law distribution is at least a statistically plausible model of the networks' degree distributions. In the Weak case, this power-law scaling covers at least 50 nodes, a relatively modest requirement. These two categories account for only 29% and 19% of networks, respectively, indicating that it is uncommon for a network to exhibit direct statistical evidence of scale-free degree distributions.

Finally, only 10% and 4% of network data sets can be classified as belonging to the Strong or Strongest categories, respectively, in which the power-law distribution is not only statistically plausible, but the exponent falls within the special $\alpha \in (2, 3)$ range and the power law is a better model of the degrees than alternatives. Taken together, these results indicate that genuinely scale-free networks are far less common than suggested by the literature, and that scale-free structure is not an empirically universal pattern.

The balance of evidence for or against scale-free structure does vary by network domain

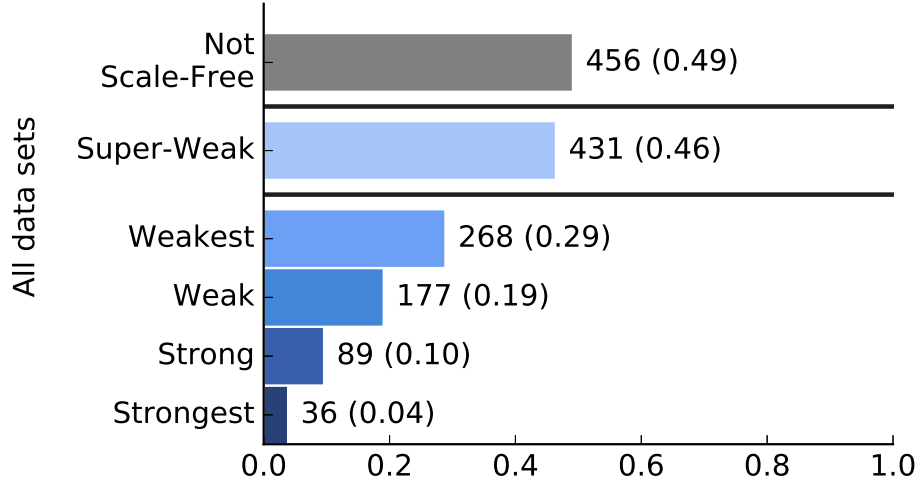


Figure 3.7: Proportion of networks by scale-free evidence category. Bars separate the Super-Weak category from the nested definitions, and from the Not Scale Free category, defined as networks that are neither Weakest or Super-Weak.

(Fig. 3.8). These variations provide a means to check the robustness of our results, and can inform future efforts to develop new structural mechanisms. We focus our domain-specific analysis on networks from biological, social, and technological sources (91% of the corpus).

Among biological networks, a majority lack any direct or indirect evidence of scale-free structure (63% Not Scale Free; Fig. 3.8a), in agreement with past work on smaller corpora of biological networks [56]. The aforementioned fungal networks represent a large share of these Not Scale Free networks, but this group also includes some protein interaction networks and some food webs. Among the remaining networks, one third exhibit only indirect evidence (33% Super-Weak), and a modest fraction exhibit the weakest form of direct evidence (19% Weakest). This latter group includes cat and rat brain connectomes. Compared to the corpus as a whole, biological networks are slightly more likely to exhibit the strongest level of direct evidence of scale-free structure (6% Strongest), and these are primarily metabolic networks.

We note that the fungal networks comprise 28% of the corpus and our analysis places 100% of them in the Not Scale Free category. Given their spatially embedded nature, it could be argued that these networks were unlikely to be scale-free in the first place. Because we know *a posteriori*

that these networks are Not Scale Free, omitting them will necessarily increase the fraction of networks in at least some of the other categories. We find that these increases occur primarily in the weaker evidence categories: 5% of non-fungal networks fall into the Strongest category (up from 4%), 13% in Strong (from 10%), 27% in Weak (from 19%), 40% in Weakest (from 29%), and 65% Super-Weak (from 46%). Hence, the qualitative conclusions from our primary analysis are robust to the inclusion of this particular subset of networks.

In contrast, social networks present a different picture. Like the corpus overall, half of social networks lack any direct or indirect evidence of scale-free structure (50% Not Scale Free; Fig. 3.8b), while indirect evidence is slightly less prevalent (41% Super-Weak). The former group includes the Facebook100 online social networks, and the latter includes many Norwegian board of director networks.

However, among the categories representing direct evidence of scale-free structure, more networks fall into the Weakest (48%) and Weak (31%) categories, but not a single network falls into the Strong or Strongest categories. Hence social networks are at best only weakly scale free, and even in cases where the power-law distribution is plausible, non-scale-free distributions are often a better description of the data. The social networks exhibiting weak evidence include many scientific collaboration networks and roughly half of the Norwegian board of director networks.

Technological networks exhibit the smallest share of networks for which there is no evidence, direct or indirect, of scale-free structure (8% Not Scale Free; Fig. 3.8c), and the largest share exhibiting indirect evidence (90% Super-Weak). The former group includes some digital circuit networks and various water distribution networks. Among the categories representing direct evidence, less than half exhibit the weakest form of direct evidence (42% Weakest). This group includes roughly half of CAIDA's networks of autonomous systems, several digital circuit networks, and several peer-to-peer networks. In contrast to biological or social networks, however, technological networks exhibit a modest fraction with strong direct evidence of scale-free structure (28% Strong). Networks in this category include the other half of the CAIDA graphs. But, almost none of the technological networks exhibit the strongest level of direct evidence (1% Strongest).

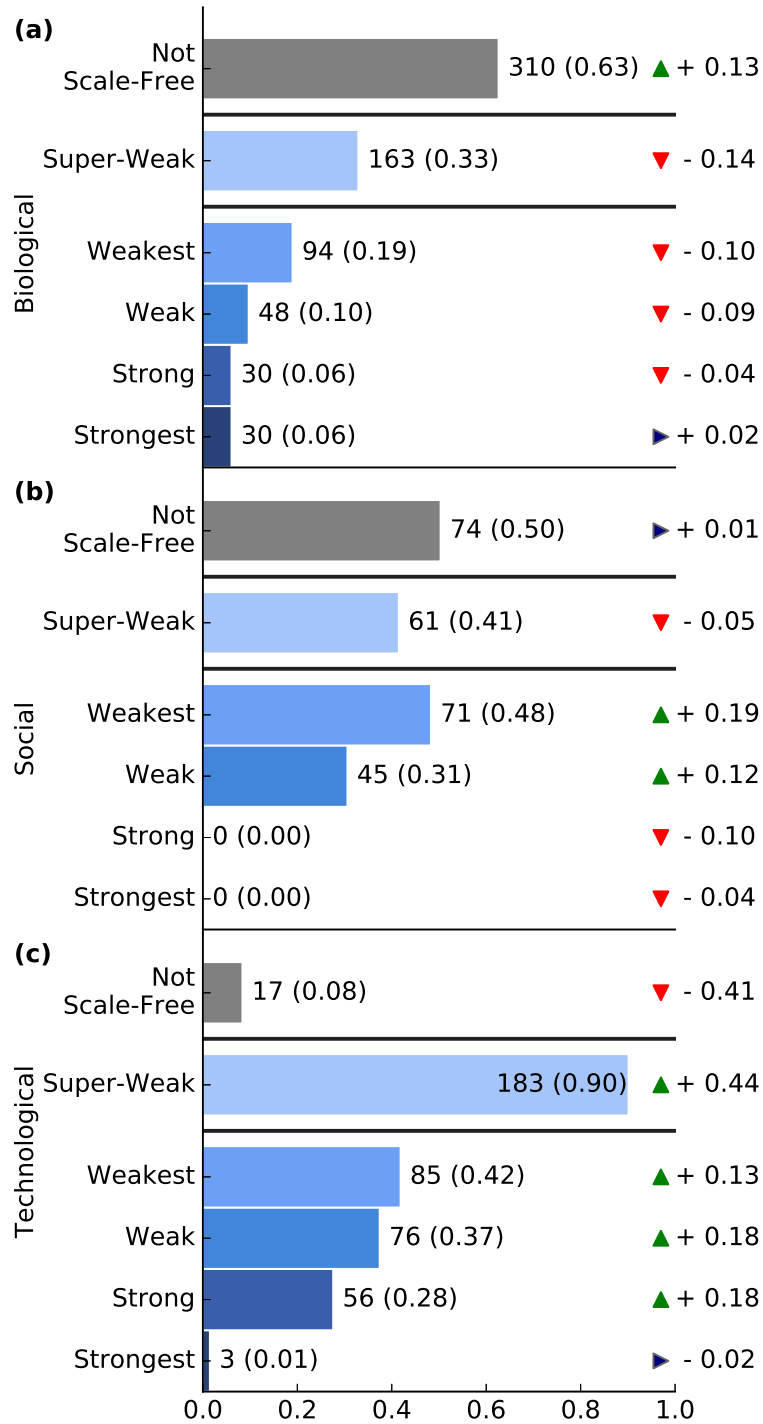


Figure 3.8: Proportion of networks by scale-free evidence category and by domain. (a) Biological networks, (b) social networks, and (c) technological networks. Tickers show change in percent from the pattern in all of the data sets.

Transportation networks do not represent a large enough fraction of the corpus for a similar statistical analysis, but do offer some useful insights for future work. Most of these networks exhibit little evidence of scale-free structure. For example, all three airport networks and 46 of 49 road networks fall into the Not Scale Free category, while two of the remaining three road networks fall into the Weak category and one into Super-Weak. All of the subway networks fall into the Super-Weak category, and nearly all fall into the Weakest category. These results suggest that scale-free networks may represent poor models of many transportation systems.

Although these results are compelling, a number of questions remain as to the robustness of these results to variations in the evaluation. In the next Chapter, we perform a series of tests of the results by modifying the methods, scale-free definitions, and looking at subsets of the corpus. This allows us to validate the results and assess their reliability.

Chapter 4

Robustness of results for power-law patterns in networks¹

The methodology we use in Chapter 3 may introduce a variety of biases into the results and we wish to rule this out. In order to assess the dependence of these results on the evaluation scheme itself, we conduct a series of robustness tests.

Specifically, we test whether the results hold qualitatively when (i) we consider only network data sets that are naturally simple (unweighted, undirected, monoplex, and no multi-edges); (ii) we remove the power-law with cutoff from the set of alternative distributions; (iii) we lower the percentage thresholds for all categories to allow admission if any one constituent simple graph satisfies the requirements; and (iv) we analyze the scaling behavior of the degree distribution's first and second moment ratio. Details for each of these tests, and two others, are given in Section 4.1. We also test whether the evaluation scheme correctly classifies four different types of synthetic networks with known structure, both scale free and non-scale free. Details and results for these tests are given in Section 4.2.

The first test evaluates whether the extension of the scale-free hypothesis to non-simple networks and the corresponding graph-simplification procedure biases the results. The second evaluates whether the presence of finite-size effects drives the lack of evidence for scale-free distributions. Applied to the corpus, each test produces qualitatively similar results as the primary evaluation scheme, indicating that the lack of empirical evidence for scale-free networks is not driven by these particular choices in the evaluation scheme itself.

¹ This chapter is adapted from: **A. D. Broido** and A. Clauset. Scale-free networks are rare. *Nature Communications* **10**: 1017 (2019)

The third considers a “most permissive” parameterization, which evaluates the impact of our requirements that a minimum percentage of degree sequences satisfy the constraints of a category. Under this test, we specifically examine how the evidence changes if we instead require that only one degree sequence satisfies the given requirements. That is, this test lowers the threshold for each category to be maximally permissive: if scale-free structure exists in any associated degree sequence, the network data set is counted as falling into the corresponding category. The fourth test provides a model-independent evaluation of a key prediction of the scale-free hypothesis.

Overall, the results of these tests corroborate our primary findings of relatively little empirical evidence for the ubiquity of scale-free networks, and suggest that empirical degree distributions exhibit a richer variety of patterns, many of which are lower variance, than predicted by the scale-free hypothesis.

4.1 Robustness checks

4.1.1 Results for simple networks alone

Extending the scale-free network hypothesis to apply to networks that are not naturally simple allowed us to draw on a much larger range of empirical network data sets. It is therefore possible that the non-simple network data sets present in the corpus have structural patterns distinct from those of simple networks, and hence are less likely to exhibit a scale-free pattern. We test for this possibility by examining the classifications of the 180 simple networks within the corpus. Among these networks, a minority exhibit neither direct nor indirect evidence of scale-free structure (53% Not Scale Free), and a modest majority exhibit at least indirect evidence (40% Super-Weak; Figure 4.1). Compared to the overall corpus, there is a notable increase in the Weakest and Weak categories. These differences can be partly explained by the distribution of simple graphs by domain, as 72% of simple graphs in the corpus are social, which exhibits similar proportions across the evidence categories. Hence, the structural diversity of real-world networks observed for the corpus as a whole is also observed when we restrict our analysis to only simple graphs, and

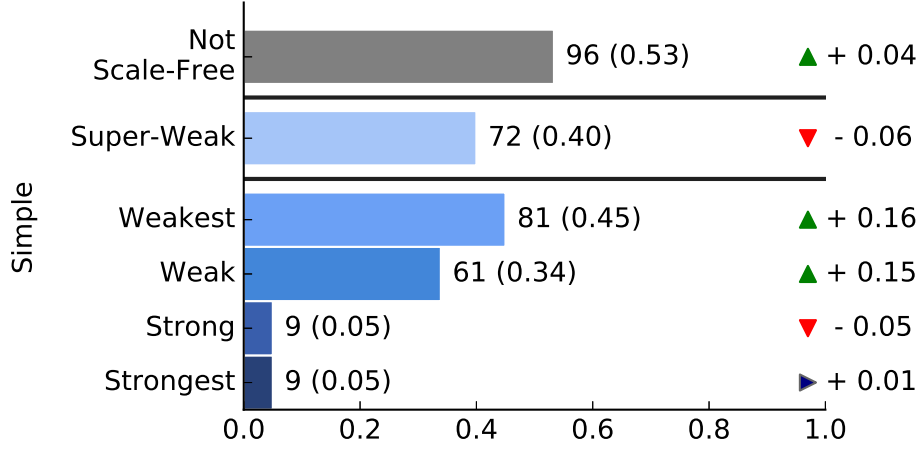


Figure 4.1: Proportions of networks in each scale-free evidence category for simple networks.

neither our inclusion of non-simple graphs, nor the graph simplification procedure described above, have skewed our results.

4.1.2 Results after removing power law with exponential cutoff from alternatives

To rule out potential bias against the scale-free hypothesis as a result of the inclusion of a power-law-like alternative in the Strong and Strongest evidence categories, we also examine the results when we remove the power law with exponential cutoff from our list of alternative distributions. As the power law is a special case of the power law with cutoff, our likelihood-ratio test can only be inconclusive or result in favor of the power law with cutoff. In the case where the power law with cutoff is the best model, this case cannot be placed in the Strongest or Strong scale-free categories by definition. In our primary evaluation, 9.59% of data sets fall into the Strong category. When we include data sets for which the power law with exponential cutoff was favored over the power law, this increases negligibly to 10.4% of data sets.

Additionally, if we also remove the restriction on the range of $\hat{\alpha}$, the percentage of data sets in this Strong category increases to 18%. This is very close to the results for the Weak category (19%), which indicates that the majority of the decrease from the Weak to the Strong is due to the imposition of the bounds on $\hat{\alpha}$ rather than the requirement against favoring alternative

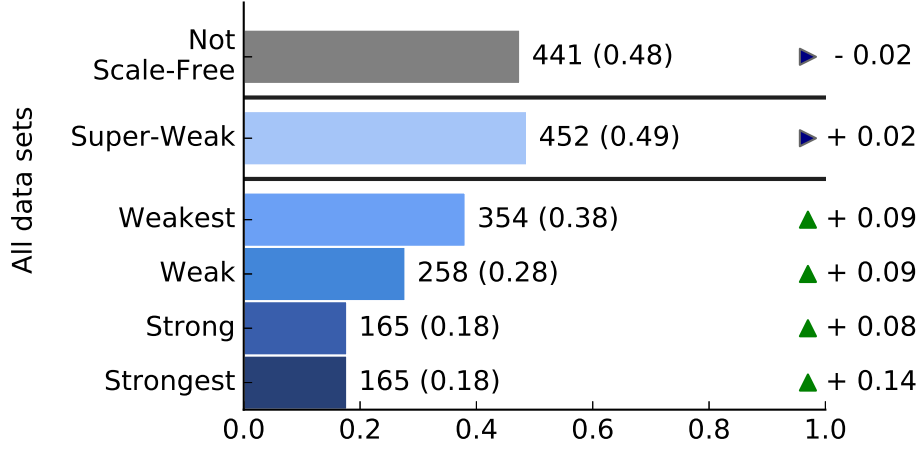


Figure 4.2: Proportions of networks in each scale-free evidence category with removed degree percentage requirements.

distributions.

There is a similarly negligible increase in the number of data sets in the Strongest category, from 3.88% to 4.63% when we allow data sets for which the power law with exponential cutoff is favored. This shift is consistent with the fact that the construction of our likelihood ratio test favors the power-law distribution since all alternatives inherit the k_{\min} that maximizes the likelihood of the power-law fit, rather than choosing their own best-fitting value.

4.1.3 Results after removing percent constraints

To rule out biases resulting from testing combining results for multiple degree sequences for each network, we adjust our scale-free categories to entirely remove percentage constraints. A network is classified as a particular type of scale free if at least one of its corresponding degree sequences satisfied the definition of that type. Under this modification, in which the Strong and Strongest categories become equivalent, and 18% of network data sets fall into this combined category (Fig. 4.2). We note that under this modified evaluation, synthetic directed networks assembled by preferential attachment should and do fall into the Strongest category of evidence. The most permissive category, Super-Weak, only changes slightly from 46% to 49%.

Because directed networks are often a specific focus within the scale-free literature, we also

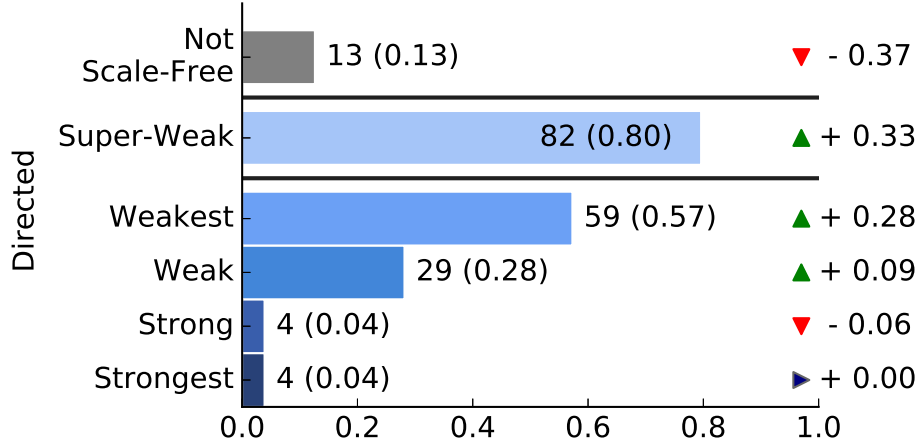


Figure 4.3: Proportions of networks in each scale-free evidence category for directed networks with removed degree percentage requirements.

examine the results for the 103 directed networks in our corpus, under the “maximally permissive” alternative parameterization of the evidence categories. In this parameterization, a directed network with a power-law distribution in the in-degrees should be and is classified as Strongest. We find that the distribution of these data sets across the evidence categories (Figure 4.3) is very close to the results over the entire corpus, implying that our evaluation scheme is not biased against directed scale-free networks. Together these tests demonstrate that the percentage requirements used in the category definitions of the primary evaluation scheme are not overly restrictive, and our qualitative conclusions are robust to variations in the precise thresholds the evaluation uses.

4.1.4 Results for the largest connected components alone

The graph simplification process described above, and used in the primary evaluation, considers all components in a given graph. As an alternative specification, we consider a check for connectedness of a network: If a network is not connected (i.e., it contains more than one component), we extract two degree sequences, one for the largest connected component, and one for the entire graph.

Including degree sequences for each largest connected in a network data set produces quantitatively similar results as when excluding it, and the overall conclusions remain unchanged. The

proportion of networks in each scale-free category differs by at most 6% from the results in Chapter 3.

4.1.5 Results for scaling behavior of degree heterogeneity

Scale-free distributions are mathematically unusual because only the moments $\langle k^m \rangle$ for $m < \alpha - 1$ are finite, and all higher moments diverge [74], asymptotically. Hence, in the most widely analyzed range of $\alpha \in (2, 3)$ for scale-free networks, the moment ratio $\langle k^2 \rangle / \langle k \rangle^2$ diverges as the network size n increases. This behavior underpins the practical relevance of many theoretical analyses of scale-free networks. Of course, diverging moments cannot be identified from finite-sized networks, and no real-world network can validate this prediction of the scale-free hypothesis. However, if most networks are scale free in this way, the scaling behavior of their moment ratios should exhibit a strongly diverging trend. Across the corpus as a whole, we find little evidence of a general pattern of diverging moment ratios (Fig. 4.4). Instead, we find enormous variation in ratios across networks, domains, and scales, such that networks with $10^2 \leq n \leq 10^3$ often have larger ratios than networks several orders of magnitude larger, and even those moments that do appear to increase with n do not increase fast enough to be consistent with scale-free behavior. We leave a more detailed investigation of these variations for future work.

We also consider a second test using the naturally simple networks, which are characterized by a single degree sequence. Given the fitted power-law distribution for each such network, we generated synthetic networks whose degree distribution is given by a semi-parametric model: the degrees below k_{\min} are given by the empirical frequencies, while the degrees at and above k_{\min} are given by the fitted power-law distribution. Hence, these synthetic networks are scale-free networks, by construction. For each simple network in this set, we generated 12 synthetic networks and compared the degree heterogeneity statistic $\langle k^2 \rangle / \langle k \rangle^2$ as a function of n for the empirical and synthetic degree distributions.

The synthetic networks, especially at larger sizes, tend to have a larger variance than the empirical distributions (Fig. 4.5), indicating that the empirical networks have substantially less

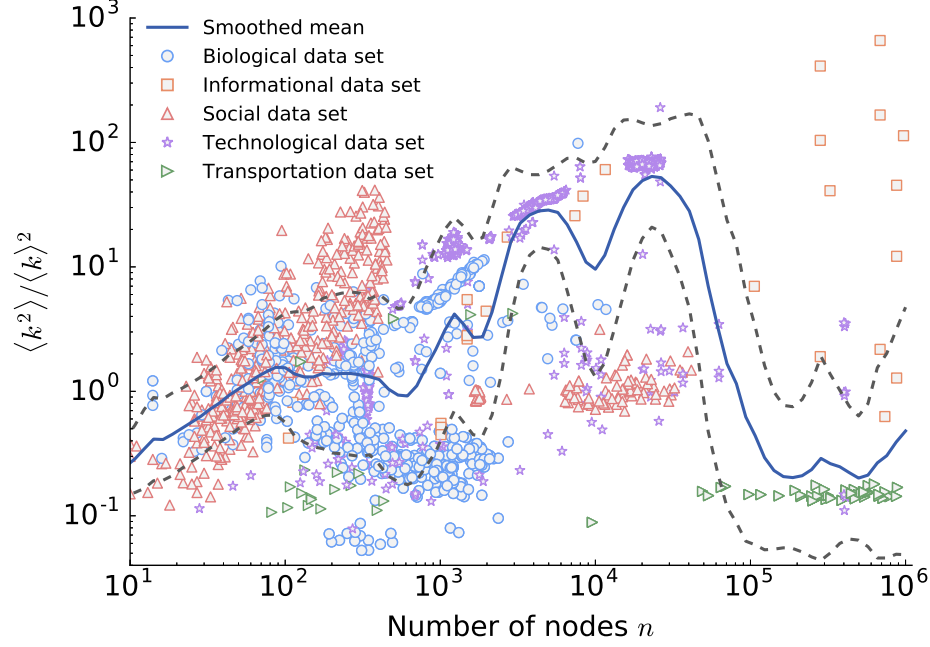


Figure 4.4: Moment ratio scaling. For 3662 degree sequences, the empirical ratio of the second to first moments $\langle k^2 \rangle / \langle k \rangle^2$ as a function of network size n , showing substantial variation across networks and domains, little evidence of the divergence pattern expected for scale-free distributions, and perhaps a roughly sublinear scaling relationship (smoothed mean via exponential kernel, with smoothed standard deviations).

degree heterogeneity than would be predicted if they were, in fact, scale free. That is, the scaling of these empirical moment ratios is not diverging as quickly as predicted by the scale-free hypothesis.

4.1.6 Results of model comparisons using information criteria

Information criteria are a common approach for selecting the best model from among a set of fitted models [20]. As an alternative to the normalized likelihood ratio test approach we use in our primary evaluation scheme, we now describe and apply an alternative model comparison method based on replacing the LRT with an application of the Akaike information criterion (AIC).

Under the AIC, a model’s adjusted “score” is written as $2k - 2 \log \mathcal{L}$, where k is the number of model parameters and \mathcal{L} is the model’s likelihood when fitted to the data. The power-law distribution used here is considered to have two estimated parameters: one in the form of α , the scaling exponent, and one in the form of the minimum value k_{\min} , which determines the left truncation

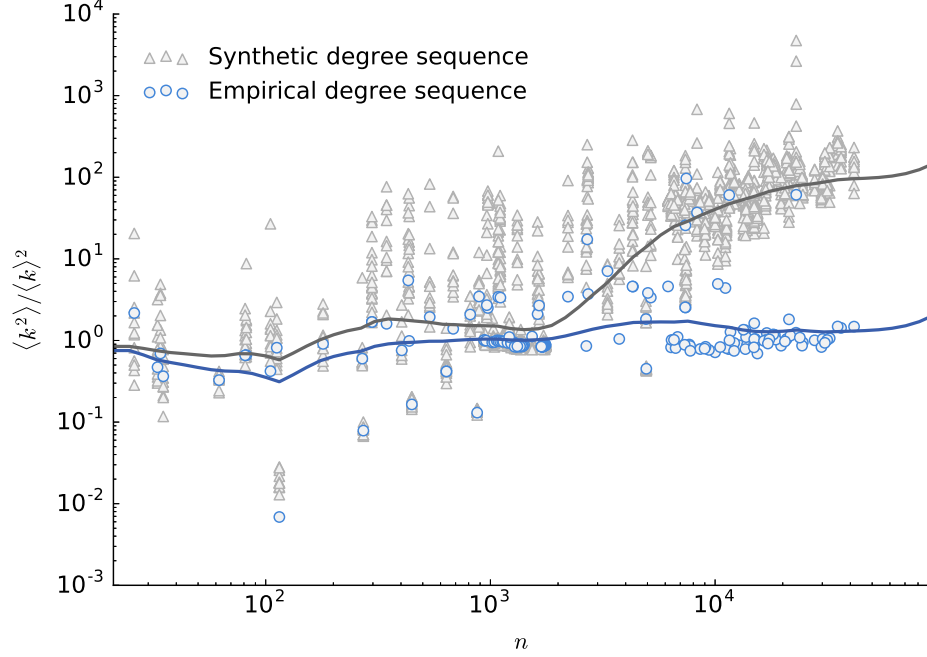


Figure 4.5: Scatterplot of the degree heterogeneity factor for empirical and synthetic simple networks vs their size. Blue points are empirical networks and 12 synthetic networks were generated from the best power-law fit for each, shown in grey.

of the degree sequence to be fitted. Because all alternative distributions in our comparison inherit the value of k_{\min} from the fitted power law, this minimum value is not considered a parameter for them. Hence, all alternative distributions have exactly two parameters, except for the exponential, which has one.

The Bayesian information criterion (BIC) (sometimes called the Schwarz criterion) is another commonly used method to compare models, but it offers little utility over the AIC in the particular setting considered here. The BIC score is written as $k \log n - 2 \log \mathcal{L}$, where n is the number of observations fitted by the model. Hence, the BIC imposes a stronger, sample-size-dependent separation between models with different complexities (number of parameters) compared to AIC. However, because all distribution models considered in our evaluation have exactly two parameters, except for the exponential which has one, the BIC will offer little insight beyond what is already provided by the AIC. For this reason, we focus our analysis on the AIC and mention results for the BIC when relevant.

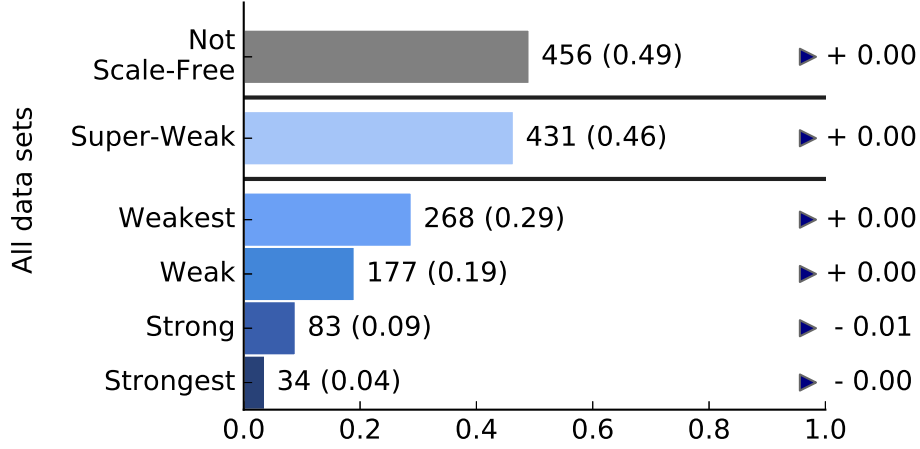


Figure 4.6: Proportions of networks in each scale-free evidence category using AIC instead of LRT for comparison of alternative distributions. Tickers indicate percent change from the results in Chapter 3.

Alternative	$p(x) \propto f(x)$	Test Outcome		
		M_{PL}	Inconclusive	M_{Alt}
Exponential	$e^{-\lambda x}$	36%	13%	51%
Log-normal	$\frac{1}{x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$	14%	31%	55%
Weibull	$e^{-\left(\frac{x}{b}\right)^a}$	37%	13%	50%
Power law with cutoff	$x^{-\alpha} e^{-\lambda x}$	0	42%	58%

Table 4.1: Comparison of scale-free and alternative distributions, using AIC. The percentage of network data sets that favor the power-law model M_{PL} , alternative model M_{Alt} , or neither, under a standard AIC comparison (see text), along with the form of the alternative distribution $f(x)$.

For each degree sequence, we compare the power-law model’s AIC score with the AIC score of each alternative distribution, deriving ΔAIC . Following standard practice, if $\Delta\text{AIC} < 2$, we conclude that there is little or no statistical evidence that the models fit the data differently [17]. In this case, we say that the comparison is inconclusive and cannot distinguish between the two models. (This outcome is comparable to failing to reject the null of $\mathcal{R} = 0$ in the normalized LRT.) Otherwise, when $\Delta\text{AIC} \geq 2$, we conclude that the model with the lower AIC value provides the better fit to the data.

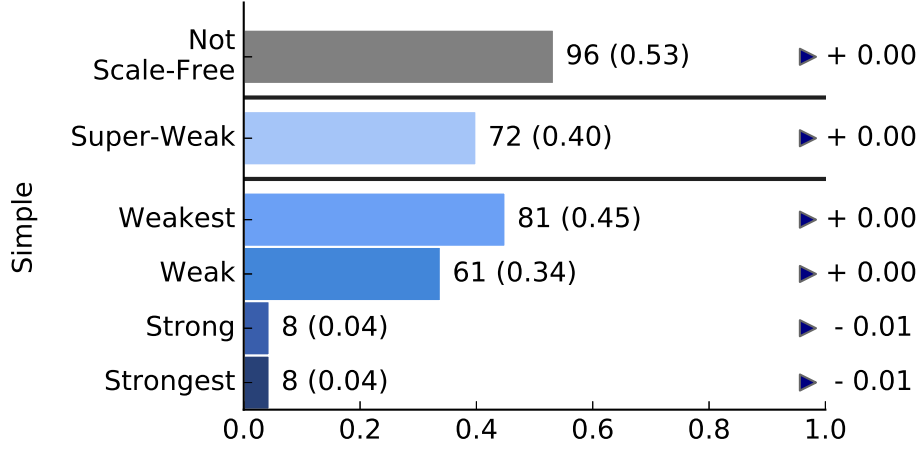


Figure 4.7: Proportions of simple networks in each scale-free evidence category using AIC instead of LRT for comparison of alternative distributions. Tickers indicate percent change from the results for simple networks in Chapter 3.

Under the AIC approach to comparing models, the percentages of network data sets that either favor the power-law model, favor the alternative model, or are inconclusive (Table 4.1) are very close to those produced under the normalized LRT used in the primary evaluation. In fact, we note that the results are slightly more more in favor of each alternative distribution under the AIC than under the LRT. Using the BIC instead of the AIC produces identical percentages for all distributions except the exponential, as explained above. The BIC results favor the exponential distribution more strongly than the AIC, in which only 16% of data sets favor the power-law model under the BIC, while 77% favor the exponential. For categorizing data sets according to their levels of evidence for scale-free structure, we only used the AIC below, as using the BIC would not change our conclusions.

In order to use an information criterion to make the model comparisons necessary to categorize a data set, we replace the LRT comparison with an AIC-based comparison, following the AIC rules stated in the preceding paragraph for concluding whether one distribution or another is favored. In this way, the category definitions themselves, and hence their interpretation, do not change, and we have only changed the method by which we decide whether an alternative distribution is favored over the power law. For succinctness, we repeat, without modification, the text

of those definitions here:

Super-Weak	For at least 50% of graphs, no alternative distribution is favored over the power law.
Weakest	For at least 50% of graphs, a power-law distribution cannot be rejected ($p \geq 0.1$).
Weak	Requirements of <i>Weakest</i> , and the power-law region contains at least 50 nodes ($n_{\text{tail}} \geq 50$).
Strong	Requirements of <i>Weak</i> and <i>Super-Weak</i> , and $2 < \hat{\alpha} < 3$ for at least 50% of graphs.
Strongest	Requirements of <i>Strong</i> for at least 90% of graphs, and requirements of Super-Weak for at least 95% of graphs.
Not Scale-Free	Networks that are neither Super-Weak nor Weakest.

We note that the percentage thresholds given in the Strongest category were chosen to match the expected error rates of the LRT. While there is no equivalent expectation for the AIC, we retain these thresholds for the sake of consistency and ease of comparison with the results of our primary evaluation.

Under this AIC-based evaluation, we find that the proportion of networks in each scale-free evidence category is nearly identical to the results produced using likelihood ratio tests (Fig. 4.6). This robustness indicates that our conclusions are not driven by the assumptions of the particular method by which we compare alternative distributions to the power-law model. Moreover, applying the AIC-based evaluation to only the simple networks, as a further robustness check, produces nearly identical results to that of using the likelihood ratio tests (Fig. 4.7), again indicating that our conclusions are robust to variations in how models are compared.

4.2 Evaluating the method on synthetic data with ground truth

To test the accuracy of the fitting, comparing, and testing methods, and the overall evaluation scheme itself, we ran four types of synthetic network data sets with known ground truth structure through the analysis pipeline. For each type, we conducted a numerical experiment using 100 instances of $n = 5000$ node network data sets. Three of these types generate scale-free structure by design: (i) one generated by a simple version of linear preferential attachment [34], (ii) one by a simple vertex copying model [78], and (iii) one by the configuration model [37] to create a temporal network where every snapshot is scale free (with $n = 1000$ nodes). The fourth type generates non-scale-free networks by design, (iv) using Erdős-Rényi random graphs.

The first type of synthetic network is generated by a simple version of linear preferential attachment [34], which is one of the most commonly referenced mechanisms for generating scale-free networks. The process is as follows, and results in a directed, unweighted, connected network. The assembly process begins with a $n = 4$ node directed network, in which each node has $k^{(\text{out})} = 3$ out edges, one to each of the other nodes. We then add one node at a time until we reach a total of $n = 5000$ nodes in the network. Each added node forms $k^{(\text{out})} = 3$ out edges. For each out edge, with probability $p = 2/3$ the connection is formed preferentially, i.e., the new node i connects to an existing node j with probability proportional to j 's in-degree $k_j^{(\text{in})}$. Otherwise, the connection is formed uniformly, i.e., the new node i connects to an existing node j with constant probability. The in-degrees distribution of the final network is scale free, following a power law of the form $k^{-2.5}$, while the out-degree distribution is a delta function at $k = 3$. The graph simplification procedure takes this directed network and produces three degree sequences, corresponding to the in-, out-, and total degrees. The in- and total degree sequences have power-law tails (the total degree sequence follows a power law for $k \gg 3$). Hence, we would expect these networks to fall into the Strong category because 2 of the 3 degree sequences are scale free.

Under our primary evaluation scheme, with thresholds set as described in Chapter 3, we find that 89% of the synthetic networks assembled by this simple model of linear preferential attachment

fall into the Super-Weak category. Omitting the power law with cutoff as an alternative model increases this rate to 97%, meaning that only 3% of the time, some alternative is a better fit to the data than is a scale-free distribution. Considering the plausibility of the fitted power laws, we find that 54% of these networks fall into the Weakest and Weak categories, 52% in the Strong category, and none in the Strongest category. As expected, the in-degree sequences and total degree sequences are generally plausible power laws (80% and 67%, respectively), while the out-degree sequences never are. The modest deviations of the plausibility rates for the in- and total degree sequences from the expected rate of 90% (which is set by the choice of critical threshold for the null hypothesis test) are likely attributable to finite-size effects.

The absence of these networks in the Strongest category is entirely due to the fact that this category requires that 90% of associated simple graphs be plausibly power law, while theoretically, only 67% (2 of 3) of the simple graphs can be. While it may seem counter-intuitive to some that preferential attachment networks, a canonical example of a scale-free network in the literature, do not fall into the Strongest category, this result is by construction because every associated degree sequence is given an equal weight in the classification scheme. However, under the maximally permissive parameterization of the evaluation scheme, in which we relax the threshold requirements to allow inclusion in a category if even one degree sequence meets the requirements, i.e., if either the in-, out-, or total degree sequences are plausibly scale free, then 93% of preferential attachment networks fall into the Strongest category.

The second type of synthetic network is generated by a simple vertex-copying model [78], and also produces scale-free structure. The process is as follows, and results in an directed, unweighted, connected network. The assembly process begins with a $n = 4$ node directed network, in which each node has $k^{(\text{out})} = 3$ out edges, one to each of the other nodes. We then add one node at a time until we reach a total of $n = 5000$ nodes in the network. For each new node v we add, we first pick an existing node u uniformly at random. Then, for each edge (u, w) , we add an edge (v, w) with probability $q = 0.6$, i.e., v copies u 's link to w . Otherwise, we choose a uniformly random

node x and add the edge (v, x) , i.e., v choose a uniformly random node to link to. This process is repeated for each of the $k_v^{(\text{out})} = 3$ outgoing edges u has. The in-degree distributions of the final network is scale free, following a power law of the form $k^{-\alpha}$, with $\alpha = 1 + \frac{1}{q} = 2.67$, while the out-degree distribution is a delta function at $k = 3$. The graph simplification procedure takes this directed network and produces three degree sequences, corresponding to the in-, out-, and total degrees. The total-degree distribution looks like $k^{(\text{in})} + k^{(\text{out})} = k^{-2.67} + 3 \approx k^{-2.67}$ for $k \gg 3$. Hence, we would expect these networks to fall into the Strong category because 2 of the 3 degree sequences are scale free.

Under the primary evaluation scheme, with thresholds set as described in Chapter 3, we find that 83% of these synthetic networks graphs fall into the Super-Weak category. Omitting the power law with cutoff as an alternative increases this rate to 97%. Furthermore, we find that 72% fall into the Weakest and Weak categories, meaning the power law is plausible with at least 50 points in the tail of the degree sequence, and 68% fall into the Strong category and none in the Strongest category. Because only 2 of the 3 degree distributions have power-law tails, the same reasoning for preferential attachment networks applies here. And, under the maximally permissive parameterization of the evaluation scheme, we find that 97% of these networks fall into the Strongest category.

The third type of synthetic network is generated using the configuration model [37], and produces a network that is expected to fall into the Strongest category, i.e., a network where every associated degree sequence is scale free. Toward this end, we construct a temporal network, where each of $T = 20$ layers has a degree sequence of $n = 1000$ nodes drawn iid from a power-law distribution with $\alpha = 2.5$. To connect the nodes in a given layer, we use the Havel-Hakimi algorithm [45, 43] to generate an initial condition for a degree-preserving edge-swapping algorithm that can sample uniformly at random from the set of simple graphs with the specified degree sequence [37].

Under the primary evaluation scheme, with thresholds set as described in Chapter 3, we

find that 100% of these synthetic networks fall into the Super-Weak, Weakest, Weak, and Strong categories, and 59% fall into the Strongest category. This latter rate falls below the expected rate, likely because of finite-size effects. Under the the maximally permissive parameterization of the evaluation scheme, 100% of these networks fall into the Strongest category.

The fourth type of synthetic network is a simple Erdős-Rényi random graph $G(n, p)$, which has no scale-free structure. In these networks, each edge exists iid with probability $p = c/(n - 1)$, where c is the mean degree. To ensure that these networks are sparse and are largely connected, we set $c = 6$. For this choice, the degree distribution is Poisson with mean c , which has a “thin” or light tail, compared to the power law.

Under the primary evaluation scheme, with thresholds set as described in Chapter 3, we find that only 15% are classified as even Super-Weak, although this rate increases to 26% if the power law with cutoff is omitted as an alternative. Furthermore, we find that 42% and 40% of these networks fall into the Weakest and Weak categories, respectively. The fitted power-law distributions for these networks all have very large scaling parameters (the smallest is $\hat{\alpha} = 6.36$), reflecting the thin-tailed structure of their degree distributions, and hence none are classified as falling into the Strong or Strongest categories. This behavior highlights the fact that a network falling into the Weakest or Weak categories can be indicative of the power-law estimation routines finding some marginal part of the extreme upper tail that is plausibly power-law distributed, even when the underlying distribution is not scale free. As $G(n, p)$ random graphs are simple, the above results are unchanged under the maximally permissive parameterization of the evaluation scheme.

4.3 Discussion

By evaluating the degree distributions of nearly 1000 real-world networks from a wide range of scientific domains, we find that scale-free networks are not ubiquitous. Fewer than 36 networks (4%) exhibit the strongest level of evidence for scale-free structure, in which every degree distribution associated with a network is convincingly scale free. Only 29% of networks exhibit the weakest

form, in which a power law is simply a statistically plausible model of some portion of the degree distribution’s upper tail. And, for 46% of networks, the power-law form is not necessarily itself a good model of the degree distribution, but is simply a statistically better model than alternatives. Nearly half (49%) of networks show no evidence, direct or indirect, of scale-free structure, and in 88% of networks, a log-normal fits the degree distribution as well as or better than a power law. These results demonstrate that scale-free networks are not a ubiquitous phenomenon, and suggest that their use as a starting point for modeling and analyzing the structure of real networks is not empirically well grounded.

Across different scientific domains, the evidence for scale-free structure is generally weak, but varies somewhat in interesting ways. These differences provide hints as to where scale-free structure may genuinely occur. For instance, our evidence indicates that scale-free patterns are more likely to be found in certain kinds of biological and technological networks. These findings corroborate theoretical work on domain-specific mechanisms for generating scale-free structure, e.g., in biological networks via the well-established duplication-mutation model for molecular networks [82, 67, 63] or in certain kinds of technological networks via highly optimized tolerance [19, 76].

In contrast, we find that social networks are at best weakly scale free, and although a power-law distribution can be a statistically plausible model for these networks, it is often not a better model than a non-scale-free distribution. Class imbalance in the corpus precludes broad conclusions about the prevalence of scale-free structure in information or transportation networks. However, the few of these in the corpus provide little indication that they would exhibit strongly different structural patterns than the better represented domains.

The variation of evidence across social, biological, and technological domains (Fig. 3.8) is consistent with a general conclusion that no single universal mechanism explains the wide diversity of degree structures found in real-world networks. The failure to find broad evidence for scale-free patterns in the degree distributions of networks indicates that much remains unknown about how network structure varies across different domains [52] and what kinds of structural patterns are common across them. We look forward to new investigations of statistical differences and

commonalities, which seem likely to generate new insights about the structure of complex systems.

The statistical evaluation here considers only the degree distributions of networks, and hence says relatively little about other structural patterns or the underlying processes that govern the form of any particular network. However, the finding that scale-free networks are empirically uncommon does imply a generally limited role for any mechanism that necessarily produces power-law degree distributions [11, 70, 91, 60], especially in domains where the evidence for strongly scale-free networks is weak, e.g., social networks. The mechanisms that govern the shape of a particular network generally cannot be determined from a static network's degree distribution alone, as it is both a weak constraint on network structure [8] and a weak discriminator between mechanisms [67]. For some networks, there is strong evidence that mechanisms like preferential attachment apply, e.g., scientific citation networks [86, 97, 91, 79]. However, the results described here imply that if such mechanisms apply more broadly, they are heavily modified or even dominated by other, perhaps domain-specific mechanisms. A claim that some network is scale free should thus be established using a severe statistical test [66] that goes beyond static degree distributions.

In theoretical network science, assuming a power law for a random graph's degree distribution can simplify mathematical analyses, and a power law can be a useful conceptual model for building intuition about the impact of extreme degree heterogeneity. And, for some types of calculations, e.g., the location of the epidemic threshold, scale-free networks can be useful models, even when real-world degree distributions are simply heavy tailed, rather than scale free [13, 104, 85]. On the other hand, if a mathematical result depends strongly on the asymptotic behavior of a scale-free degree distribution, the results' practical relevance will necessarily depend on the empirical prevalence of scale-free structures, which we show to be uncommon or rare, depending on the kind of scale-free structure of interest. Mathematical results based on extreme degree heterogeneity may, in fact, have more narrow applicability than previously believed, given the lack of evidence that empirical moment ratios diverge as quickly as those results typically assume (Fig. 4.4 and Fig. 4.5).

The structural diversity of real-world networks uncovered here presents both a puzzle and an opportunity. The strong focus in the scientific literature on explaining and exploiting scale-free

patterns has meant relatively less is known about mechanisms that produce non-scale-free structural patterns, e.g., those with degree distributions better fitted by a log-normal. Two important directions of future work will be the development and validation of novel mechanisms for generating more realistic degree structure in networks, and novel statistical techniques for identifying or untangling them given empirical data. Similarly, theoretical results concerning the behavior of dynamical processes running on top of networks, including spreading processes like epidemiological models, social influence models, or models of synchronization, may need to be reassessed in light of the genuine structural diversity of real-world networks.

The statistical methods and evidence categories developed and used in our evaluation of the scale-free hypothesis provide a quantitatively rigorous means by which to assess the degree to which some network exhibits scale-free structure. Their application to a novel network data set should enable future researchers to determine whether assuming scale-free structure is empirically justified.

Furthermore, large corpora of real-world networks, like the one used here, represent a powerful, data-driven resource by which to investigate the structural variability of real-world networks [52]. Such corpora could be used to evaluate the empirical status of many other broad claims in the networks literature, including the tendency of social networks to exhibit high clustering coefficients and positive degree assortativity [77], the prevalence of the small-world phenomena [108], the prevalence of “rich clubs” in networks [27], the ubiquity of community [39] or hierarchical structure [21], and the existence of “super-families” of networks [68]. We look forward to these investigations and the new insights they will bring to our understanding of the structure and function of networks.

Chapter 5

Comparing methods for power-law fitting

5.1 Introduction

In the previous chapters we strive to address the question of the ubiquity of scale-free structure in real-world networks. At its core, this requires fitting a power-law distribution to data. We have done so using the Kolmogorov-Smirnov (KS) method, but there are other algorithms for this fitting task. One such method is an algorithm that uses bootstrapping to find the threshold value for the power-law tail [28]. This method has the desirable property that it asymptotically minimizes the mean squared error in the estimate of the scaling parameter α , while the KS method does not have this guarantee [32]. In this chapter we consider using the bootstrap method to fit a power law to data and compare its performance with that of the KS method. We broaden our discussion to consider continuous power-law distributions because that is where we find useful theoretical guarantees.

Power-law distributions appear in many of areas of science, from physical systems [49, 99] to the number of customers affected in power outages [74], or the pattern of connections in a network [22, 6, 11, 16]. They describe systems in which there is a non-negligible probability of seeing data with high values, like cities with very large population, or nodes in a network that are highly connected. Because of their broad applicability, power laws have received broad attention in the literature [22, 6, 70, 74, 71, 92, 101, 16, 106, 32].

Over the years, as a result of their widespread relevance, there have been a variety of methods introduced to detect power laws in empirical data. In particular, we often seek not just to fit a

power-law of the form $f(x) = Cx^{-\alpha}$ to the data, but we wish to find a threshold value in the data, above which the values follow a power-law distribution [32, 46, 88, 28, 22]. Recent work has analyzed the asymptotic performance of a method called the Kolmogorov-Smirnov method on data sets with different types of underlying distributions [32]. Here we compare the empirical performance of this method with a bootstrapping procedure [28, 105].

5.2 Methods

We say that a random variables x follows a power-law distribution if $f(x) \propto x^{-\alpha}$, $\alpha > 1$. Empirical distributions very often follow a power-law distribution only for values above a certain threshold [22, 32, 46, 28]. Thus in practice we often seek not only the power-law exponent α but also the cutoff value x_{\min} where the tail begins. Typically we find x_{\min} first, and then use the corresponding truncated data set to find the maximum likelihood estimate for α .

Given a data set $X = \{x_1, x_2, \dots, x_n\}$, we denote the order statistics by $x_1^{(n)} \leq x_2^{(n)} \leq \dots \leq x_n^{(n)}$. We can then use the k th largest order statistic to indicate the tail cutoff. That is, $x_{n-k-1}^{(n)} < x_{\min} \leq x_{n-k}^{(n)}$. The tail distribution function F for the data points above the threshold is then

$$1 - F(x) \propto \left(\frac{x}{x_{\min}} \right)^{-\alpha+1} \approx \left(\frac{x}{x_{n-k}^{(n)}} \right)^{-\alpha+1},$$

where we have used $x_{n-k}^{(n)}$ to approximate x_{\min} . The MLE for α is then given by the well-known Hill estimator [46]

$$\hat{\alpha} = 1 + (k+1) \left[\sum_{i=1}^k \log x_{n-i+1}^{(n)} - \log x_{n-k}^{(n)} \right].$$

Note that this is equivalent to the MLE derived in Chapter 2 (Eqn 2.1) The Hill estimator is known to be consistent [65] and asymptotically normal with rate $k^{-1/2}$ if $n \rightarrow \infty$, $k \rightarrow \infty$ and $n/k \rightarrow 0$ [25, 29]. Thus the performance of the Hill estimator for α depends strongly on our ability to accurately estimate the value of k , or equivalently x_{\min} . Here we consider two methods for the estimation of the best fit for power-law tail of a data set.

5.2.1 Kolmogorov Smirnov Method

The first method we implement is the Kolmogorov-Smirnov (KS) Method, developed by Clauset et al. [22]. The aim is to minimize the distance between the probability distribution of the observed data and the best-fit power-law distribution for data points above the current estimate of the threshold \hat{x}_{\min} . Specifically, we estimate this tail cutoff by minimizing the KS distance between the empirical distribution of the data for the observations $x_i \geq \hat{x}_{\min}$ and the best-fit power-law distribution for these values. The value of the data at which this minimum occurs is the value we select as \hat{x}_{\min} . We define the KS distance as

$$D_k = \sup_{x \geq 1} \left| \frac{1}{k+1} \sum_{i=1}^n \mathbb{1}_{(X_i \geq x_{\min})} - \left(\frac{x}{x_{\min}} \right)^{-\hat{\alpha}+1} \right|$$

where $x_{n-k}^{(n)}$ is the first order statistic that is at least equal to x_{\min} .

We take \hat{x}_{\min} to be $x_{n-k^*}^{(n)}$ where

$$k^* = \arg \min_{k \in \{2, \dots, n\}} D_k.$$

Note that in the case of a non-unique minimum, we choose the smallest [32].

5.2.2 Bootstrapping procedure

The second method we test is a bootstrap method, developed by Danielsson et al. [28]. It is designed to minimize the asymptotic mean squared error in the estimate of the exponent of the distribution function. The method calculates γ for tail data from distributions F of the form

$$1 - F(x) = L(x)x^{-1/\gamma}$$

where $\gamma > 0$ and $L(x)$ is a slowly-varying function. Slowly varying functions satisfy

$$\lim_{x \rightarrow \infty} \frac{L(ax)}{L(x)} = 1.$$

To relate this to the pure power-law, we can choose $L(x) = C > 0$ and observe that $\alpha = 1 + 1/\gamma$. Note that both procedures are thus explicitly *tail models*, but the definition of slowly varying is

inherently an asymptotic one, while the KS-minimization method is seeking a region at finite value that looks like a power law.

Given order statistics $x_1^{(n)} \leq \dots x_n^{(n)}$ for data set $X = \{x_1, x_2, \dots, x_n\}$, the Hill estimator for γ is given by

$$\gamma_n(k) = \frac{1}{k} \sum_{i=1}^k \log x_{n-i+1}^{(n)} - \log x_{n-k}^{(n)}.$$

We seek to minimize the asymptotic mean squared error (AMSE) of this estimator $\gamma_n(k)$

$$\text{AMSE}(n, k) = \text{Asy E} \left[(\gamma_n(k) - \gamma)^2 \right],$$

where γ is the true value of the parameter. We seek the value the value k^* that minimizes this error. That is,

$$k^*(n) = \arg \min_k \text{AMSE}(n, k).$$

Since the true value for γ is unknown for real data sets, we cannot directly compute this AMSE. Instead we will use bootstrapping to estimate this error. Bootstrap samples of size n will not give an estimate of error that converges to the true error, but using bootstrap resamples of size $n_1 < n$ solves this problem [44]. We draw resamples $B = \{b_1, b_2, \dots, b_{n_1}\}$ from X with replacement. This gives us order statistics $b_1^{(n_1)} \leq \dots \leq b_{n_1}^{(n_1)}$ for B , and the Hill estimator for γ based on the bootstrap resamples is defined as

$$\gamma_{n1}(k_1) = \frac{1}{k_1} \sum_{i=1}^{k_1} \log b_{n_1-i+1}^{(n_1)} - \log b_{n_1-k_1}^{(n_1)}.$$

This gives us a bootstrap estimate for the AMSE [44, 28]

$$\widehat{\text{AMSE}}(n, k) = \text{E} \left[(\gamma_n(k) - \gamma_{n1}(k_1))^2 \mid X \right].$$

This estimate, however, still contains a term with an unknown value: as we do not know the true value of k , we cannot calculate $\gamma_n(k)$. We introduce a control variate M

$$M_n(k) = \frac{1}{k} \sum_{i=1}^k \left(\log x_{n-i+1}^{(n)} - \log x_{n-k}^{(n)} \right)$$

and note that $M_n(k)/(2\gamma_n(k))$ is a consistent estimator of γ [28]. We therefore define the bootstrap estimate for the asymptotic mean squared error by replacing $\gamma_n(k)$ in the above AMSE estimate with the bootstrapped estimate $M_{n_1}(k_1)/(2\gamma_{n_1}(k_1))$ where

$$M_{n_1}(k_1) = \frac{1}{k_1} \sum_{i=1}^{k_1} \left(\log b_{n_1-i+1}^{(n_1)} - \log b_{n_1-k_1}^{(n_1)} \right).$$

Finally we define the bootstrap estimate dependent only on n_1 and k_1 as

$$Q(n_1, k_1) = \mathbb{E} \left[\left(M_{n_1}(k_1) - 2(\gamma_{n_1}(k_1))^2 \right)^2 \mid X \right]$$

and choose k_1 to minimize this quantity.

The statistics $M_n(k)/(2\gamma_n(k)) - \gamma_n(k)$ and $\gamma_n(k) - \gamma$ both converge to zero asymptotically [28]. In fact, the k -value that minimizes $\text{AMSE}(n, k)$ and the k -value that minimizes

$$\text{Asy } \mathbb{E} \left[\left(M_{n_1}(k_1) - 2(\gamma_{n_1}(k_1))^2 \right)^2 \right]$$

are of the same general order (with respect to n), under convergence conditions [28]. The method will estimate the asymptotically optimal choice of k , and thus using this in the Hill estimator will yield an asymptotically optimal choice of γ , or in our case, α .

We choose the number of bootstrap resamples based on available computational time. All that remains is to choose n_1 . It has been shown that for any $\varepsilon \in (0, 1/2)$, setting $n_1 = n^{1-\varepsilon}$ will lead asymptotically to the optimal choice of k and γ [28]. Therefore, we use a heuristic procedure to select n_1 from this range.

It can be shown that

$$\text{Asy } \mathbb{E} \left[\left(M_n(\bar{k}(n)) - 2(\gamma_n(\bar{k}(n)))^2 \right)^2 \right] \frac{Q(n_2, k_2^*)}{(Q(n_1, k_1^*))^2} \rightarrow 1$$

in probability, where $\bar{k}(n)$ is the value of k that minimizes $\text{Asy } \mathbb{E} \left[\left(M_n(k) - 2(\gamma_n(k))^2 \right)^2 \right]$ [28]. Thus we can use the ratio

$$R(n_1) = \frac{(Q(n_1, k_1^*))^2}{Q(n_2, k_2^*)}$$

as an estimator for

$$\text{Asy } \mathbb{E} \left[\left(M_n(\bar{k}(n)) - 2(\gamma_n(\bar{k}(n)))^2 \right)^2 \right].$$

Algorithm 1 Bootstrap algorithm [28]

```

1: function FITPOWERLAW( $X$ , num_resamps)
2:    $n_1 \leftarrow \text{FINDN1}(X, \text{num\_resamps})$ 
3:    $\mathbf{B}_1 \leftarrow \text{RESAMPLE}(X, n_1, \text{num\_resamps})$ 
4:   for  $k_1 \in \text{grid}\{1, n_1, \text{step\_size}\}$  do
5:      $Q(n_1, k_1) \leftarrow \text{BOOTSTRAPMSE}(\mathbf{B}_1, k_1)$ 
6:    $k_1^* \leftarrow \arg \min_{k_1} Q(n_1, k_1)$ 
7:    $n_2 \leftarrow \text{int}(n_1^2/n)$ 
8:    $\mathbf{B}_2 \leftarrow \text{RESAMPLE}(X, n_2, \text{num\_resamps})$ 
9:   for  $k_2 \in \text{grid}\{1, n_2, \text{step\_size}\}$  do
10:     $Q(n_2, k_2) \leftarrow \text{BOOTSTRAPMSE}(\mathbf{B}_2, k_2)$ 
11:    $k_2^* \leftarrow \arg \min_{k_2} Q(n_2, k_2)$ 
12:    $\hat{k} \leftarrow \text{CALCULATEK}(Q(n_1, k_1^*), Q(n_2, k_2^*), n_1)$ 
13:    $\gamma_n(\hat{k}) \leftarrow \text{GAMMAESTIMATOR}(X, \hat{k})$ 
14:    $\alpha_n(\hat{k}) \leftarrow 1/\gamma_n(\hat{k})$ 
15:   return  $\hat{k}, \alpha_n(\hat{k})$ 

```

We choose n_1 to minimize $R(n_1)$.

Our goal is to find the power-law exponent α , rather than γ , as posed in the original algorithm. By invariance of maximum likelihood estimators, since $\alpha = 1 + 1/\gamma$, the MLE for α is $\hat{\alpha} = 1 + 1/\hat{\gamma}$. This estimate $\hat{\alpha}$ will still be the value that minimizes the asymptotic mean squared error of the Hill estimator. We present the formal algorithm and necessary helper functions below (Algorithms 1 & 2).

Algorithm 2 Helper functions for bootstrap algorithm [28]

```

1: function FINDN1( $X$ , num_resamps)
2:    $n \leftarrow \text{length}(X)$ 
3:   for  $n_1 \in \text{grid}\{n^{1/2}, n, \text{step\_size}\}$  do
4:      $R_{n_1} = \frac{Q(n_1, k_1^*)^2}{Q(n_2, k_2^*)}$ 
5:    $n_1^* \leftarrow \arg \min_{n_1} R_{n_1}$ 
6:   return  $n_1^*$ 

7: function RESAMPLE( $X$ ,  $n_1$ , num_resamps)
8:    $n \leftarrow \text{length}(X)$ 
9:   for  $i \leftarrow 1 : \text{num\_resamps}$  do
10:     $B_i \leftarrow \text{SAMPLEWITHREPLACEMENT}(X)$ 
11:    $\mathbf{B} = \{B_1, \dots, B_{\text{num\_resamps}}\}$ 
12:   return  $\mathbf{B}$ 

13: function BOOTSTRAPMSE( $\mathbf{B}$ ,  $k_1$ )
14:    $n_1 \leftarrow \text{length}(B)$ 
15:    $Q \leftarrow 0$ 
16:   for  $B_i \in \mathbf{B}$  do
17:      $B_i^{(n_1)} \leftarrow \text{sort}(B_i)$ 
18:      $M_{n_1}(k_1) \leftarrow \text{MESTIMATOR}(B_i^{(n_1)}, k_1)$ 
19:      $\gamma_{n_1}(k_1) \leftarrow \text{GAMMAESTIMATOR}(B_i^{(n_1)}, k_1)$ 
20:      $Q \leftarrow Q + (M_{n_1}(k_1) - (\gamma_{n_1}(k_1))^2)^2$ 
21:    $Q \leftarrow Q/n_1$ 
22:   return  $Q$ 

23: function CALCULATEK( $k_1^*, k_2^*, n_1$ )
24:    $\hat{k} = \frac{(k_1^*)^2}{k_2^*} \left( \frac{\log k_1^*}{2 \log n_1 - \log k_1} \right)^{2 \frac{\log n_1 - \log k_1^*}{\log n_1}}$ 
25:   return  $\hat{k}$ 

26: function MESTIMATOR( $X$ ,  $k$ )
27:    $M \leftarrow (1/k) \sum_{i=1}^k (x_{n-i+1}^{(n)} - x_{n-k}^{(n)})^2$ 
28:   return  $M$ 

29: function GAMMAESTIMATOR( $X$ ,  $k$ )
30:    $\gamma_n(k) \leftarrow (1/k) \sum_{i=1}^k x_{n-i+1}^{(n)} - x_{n-k}^{(n)}$ 
31:   return  $\gamma_n(k)$ 

```

5.3 Results

5.3.1 Continuous synthetic data

We generate continuous synthetic data to test and compare the performance of the two algorithms. We draw data from several different types of distributions and compare the performance on each. All of the distributions have power-law tails above the threshold x_{\min} , and differ in the form of the non-power-law body. The various forms of the body of the distribution affect how easy it is for the methods to detect x_{\min} accurately.

The first probability distribution is a piecewise function of the form

$$f(x) = \begin{cases} f_1(x) = C_1 e^{-\lambda x} & 1 \leq x < x_{\min} \\ f_2(x) = C_2 x^{-\alpha} & x \geq x_{\min} \end{cases}$$

so that values of x below x_{\min} are drawn from an exponential distribution, and values above are drawn from a power-law distribution. To guarantee that $f(x)$ is continuous and integrates to 1 we impose the following constraints, and to reduce the degrees of freedom of the system we require that the derivatives are related by a constant at $x = x_{\min}$:

$$(1) \quad f_1(x_{\min}) = f_2(x_{\min})$$

$$(2) \quad f_1'(x_{\min}) = \frac{1}{b} f_2'(x_{\min})$$

$$(3) \quad \int_1^{\infty} f(x) = 1.$$

By the first constraint, we have

$$C_1 e^{-\lambda x_{\min}} = C_2 x_{\min}^{-\alpha}. \quad (5.1)$$

Combining this with constraint 2 gives $\lambda = \frac{\alpha}{b x_{\min}}$, and we can rewrite Eqn. (5.1) as

$$C_1 e^{-\alpha/b} = C_2 x_{\min}^{-\alpha}. \quad (5.2)$$

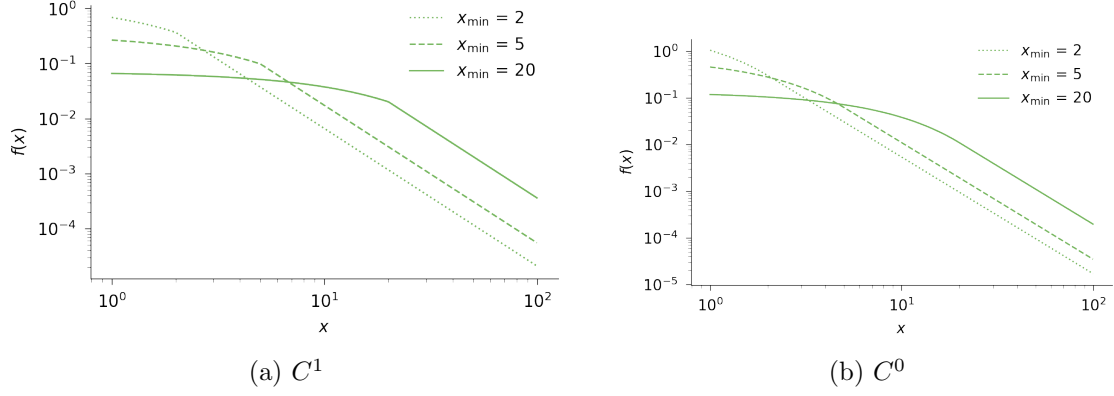


Figure 5.1: Distribution that is exponential below x_{\min} and power law above. (a) Distribution with discontinuous derivative and various x_{\min} values. Specifically, $b = 2$ (see text for details). (b) Distribution with continuous derivative (see text for details).

To find C_1 and C_2 , we integrate the normalization constraint

$$\begin{aligned}
 1 &= \int_1^{\infty} f(x) dx \\
 &= \int_1^{x_{\min}} C_1 e^{-\frac{\alpha}{b x_{\min}} x} dx + \int_{x_{\min}}^{\infty} C_2 x^{-\alpha} dx \\
 &= -\frac{b x_{\min}}{\alpha} C_1 \left[e^{-\frac{\alpha}{b}} - e^{-\frac{\alpha}{b x_{\min}}} \right] + \frac{x_{\min}}{\alpha - 1} C_2 x_{\min}^{-\alpha} \\
 &= -\frac{b x_{\min}}{\alpha} C_1 \left[e^{-\frac{\alpha}{b}} - e^{-\frac{\alpha}{b x_{\min}}} \right] + \frac{x_{\min}}{\alpha - 1} C_1 e^{-\frac{\alpha}{b}}
 \end{aligned}$$

where the last line follows from Eqn. (5.2). This gives

$$C_1 = \frac{(\alpha - 1)\alpha}{x_{\min} \left(-b(\alpha - 1) \left(e^{-\frac{\alpha}{b}} - e^{-\frac{\alpha}{b x_{\min}}} \right) + \alpha e^{-\frac{\alpha}{b}} \right)}$$

and

$$C_2 = \frac{(\alpha - 1)\alpha x_{\min}^{\alpha-1} e^{-\frac{\alpha}{b}}}{-b(\alpha - 1) \left(e^{-\frac{\alpha}{b}} - e^{-\frac{\alpha}{b x_{\min}}} \right) + \alpha e^{-\frac{\alpha}{b}}}$$

Thus

$$f(x) = \begin{cases} f_1(x) = C_1 e^{-\frac{\alpha}{b x_{\min}} x} & 1 \leq x < x_{\min} \\ f_2(x) = C_2 x^{-\alpha} & x \geq x_{\min}. \end{cases}$$

To draw from this distribution, we use inverse transform sampling. The cdf (Fig. 5.1) is

given by

$$F(x) = \begin{cases} F_1(x) & 1 \leq x < x_{\min} \\ F_2(x) & x \geq x_{\min}, \end{cases}$$

where

$$F_1(x) = \frac{b(\alpha - 1) \left(e^{-\frac{\alpha}{bx_{\min}}} - e^{-\frac{\alpha}{b}x} \right)}{b(\alpha - 1)e^{-\frac{\alpha}{bx_{\min}}} + \alpha e^{-\frac{\alpha}{b}}(b + \alpha - \alpha b)}$$

and

$$F_2(x) = 1 + \frac{\alpha x_{\min}^{\alpha-1} e^{\frac{\alpha}{bx_{\min}}} x^{1-\alpha}}{-b(\alpha - 1) \left(e^{-\frac{\alpha}{b}} - e^{-\frac{\alpha}{b}x_{\min}} \right) + \alpha e^{-\frac{\alpha}{b}}}$$

yielding the inverses

$$F_1^{-1}(y) = \frac{bx_{\min}}{\alpha} \log \left(-\frac{be^{\frac{1+x_{\min}}{bx_{\min}}\alpha}(\alpha - 1)}{-re^{\frac{\alpha}{bx_{\min}}}(b(\alpha - 1) - \alpha) + be^{\frac{\alpha}{b}}(\alpha - 1)(r - 1)} \right)$$

and

$$F_2^{-1}(z) = \left(\frac{e^{-\frac{\alpha}{bx_{\min}}}(r - 1)x_{\min}^{1-\alpha}}{\alpha} \right)^{\frac{1}{1-\alpha}} \times \left(-b \left(e^{\frac{\alpha}{b}} - e^{\frac{\alpha}{b}} \right) (\alpha - 1) - \alpha e^{\frac{\alpha}{bx_{\min}}} \right)^{\frac{1}{1-\alpha}}$$

To use the inverse transform method we will draw $r \sim \text{Unif}(0, 1)$ and determine which of these 2 functions to use to generate a data point x . This gives us our final inverse function to draw from

$$F^{-1}(r) = \begin{cases} F_1^{-1}(r) & r < F_1(x_{\min}) \\ F_2^{-1}(r) & r \geq F_1(x_{\min}). \end{cases}$$

For the first data sets that we draw, we set $b = 2$, fixing the discontinuity in the derivative of the pdf. In comparing the two methods, there are several factors to consider. While both methods focus on estimating α , they each find an estimate for k and, relatedly, x_{\min} along the way. The first test we perform compares the \hat{x}_{\min} values estimated by each of the two methods to the true x_{\min} values used to generate synthetic data sets. We generate 250 synthetic data sets with 10000

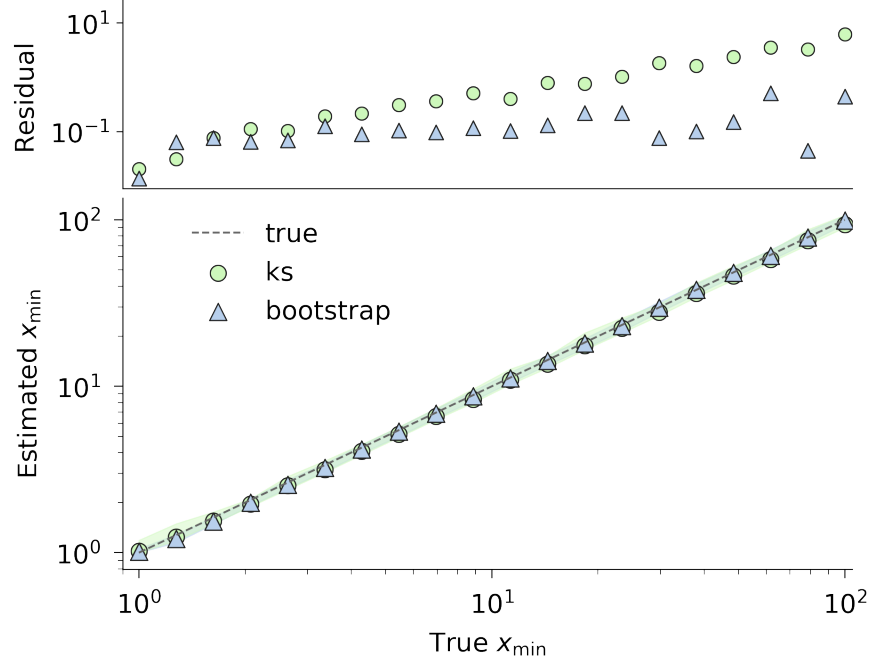


Figure 5.2: Recovery of x_{\min} for synthetic data sets with discontinuous derivative. Each data set has 10000 points, 250 at each true x_{\min} value. The points are at the median \hat{x}_{\min} over all 250 estimates and the 25th to 75th quantile range is shaded.

points for each value of x_{\min} , and then report the median estimated value along with the range from the 25th to 75th quantile. (Fig. 5.2). Both methods recover x_{\min} with very high accuracy. From the residual errors we see that the bootstrap method is slightly more accurate in estimates of x_{\min} , particularly as the value of x_{\min} increases.

We also want a sense of how well each method converges to the true value of α . We compute the RMSE of the estimate of $\hat{\alpha}$ as a function of the size of the data set, for values of n between 100 and 5000. We then compute the RMSE at each size over 100 draws from our synthetic distribution with $\alpha = 2.5$ and $x_{\min} = 5$. Based on the known asymptotic convergence rate of the Hill estimator, we expect its RMSE to decrease at the same rate of $k^{-1/2}$. In this case, that is equivalent to decreasing at a rate of $n^{-1/2}$ because of how we generate the data sets with a fixed values of x_{\min} . Both methods converge with this expected rate (Fig. 5.3). For smaller data sets the bootstrap method has much higher error, but for larger data sets, the performance is very similar, and possibly slightly better.

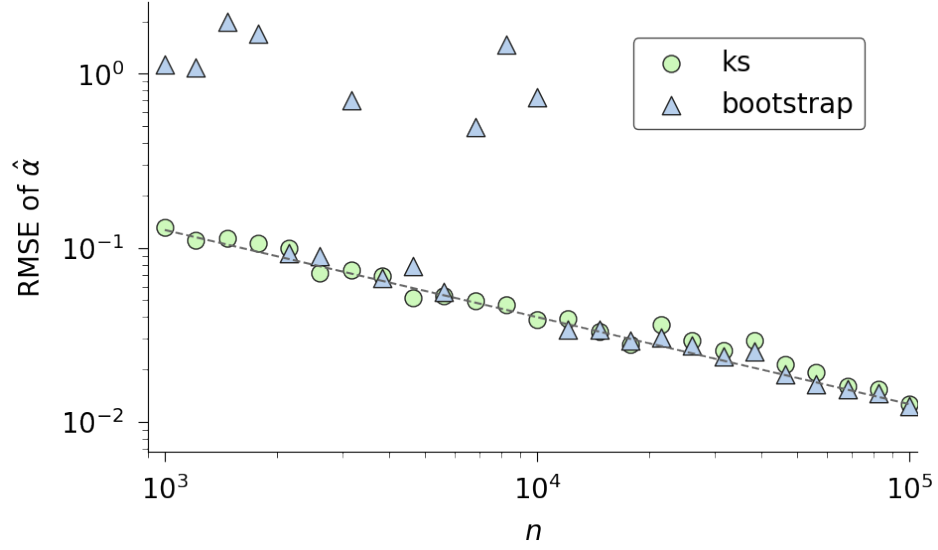


Figure 5.3: RMSE convergence comparison for bootstrap method on synthetic data sets with discontinuous derivative by number of bootstrap resamples. 100 data sets each. The dashed line is the theoretical limit $n^{-1/2}$.

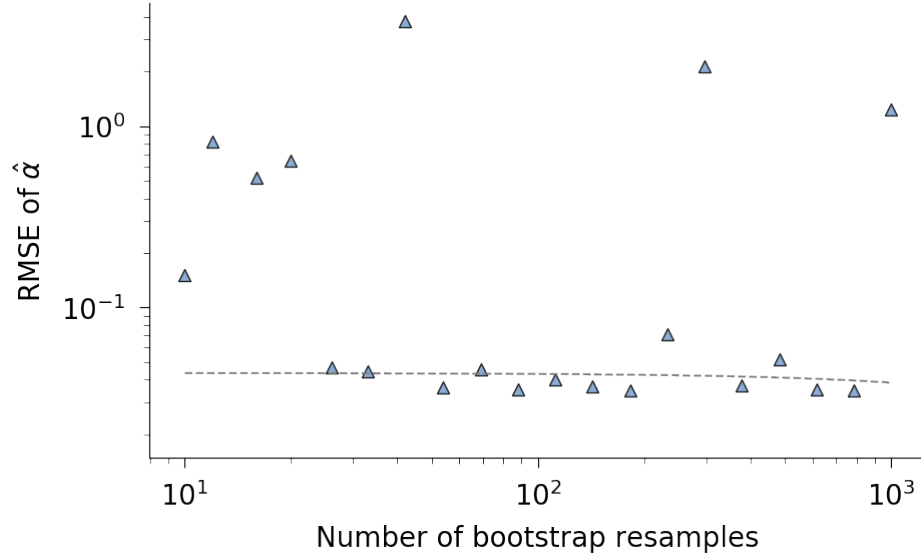


Figure 5.4: RMSE convergence comparison for bootstrap method on synthetic data sets with discontinuous derivative by number of bootstrap resamples. 250 data sets each. The dashed line is a linear regression (fit to points with $\text{RMSE} < 0.1$) with slope -4.97×10^{-6} and intercept 0.0435.

The KS method uses grid search to find the parameter value that maximizes the likelihood of the data and the threshold value that minimizes the KS distance. The bootstrap method on

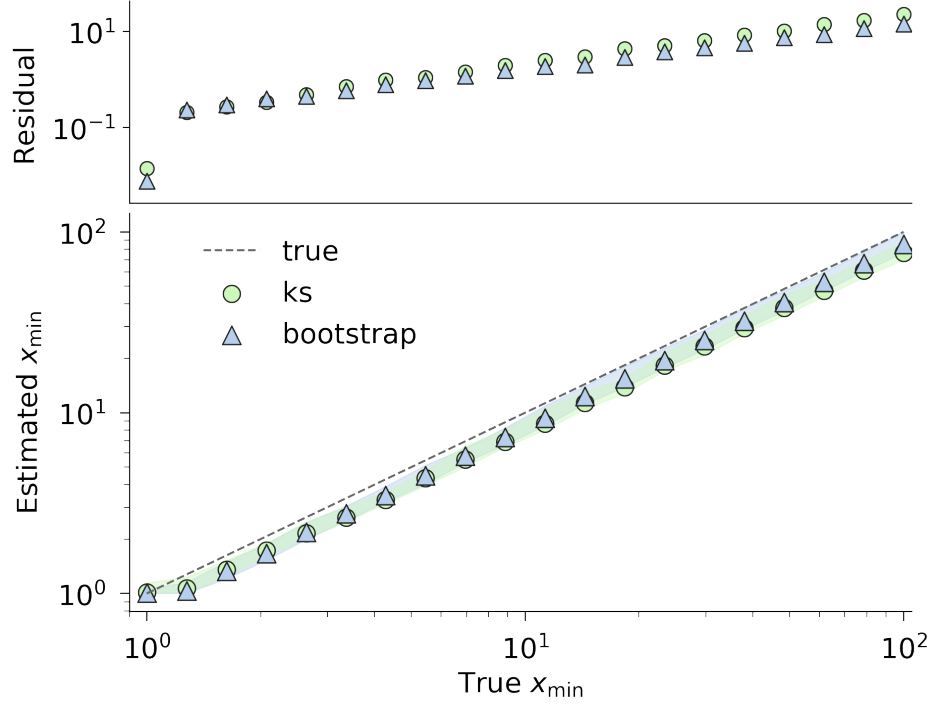


Figure 5.5: Recovery of x_{\min} for synthetic data sets with continuous derivative. Each data set has 10000 points, 100 at each x_{\min} value. The median is plotted with the 25th to 75th quantile range shaded.

the other hand has asymptotic guarantees of convergence, but no obvious metric for determining convergence of a particular call to the algorithm. The only free parameter of the algorithm is the number of bootstrap resamples so we examine the error as a function of the number of resamples. We draw 5000 synthetic datasets from the distribution with discontinuous derivative with $n = 10000$, $x_{\min} = 5$, $\alpha = 2.5$ and assess the performance of the algorithm at various values of the number of bootstrap resamples between 10 and 1000 (Fig. 5.4). We do see a slight downward trend in the RMSE of $\hat{\alpha}$, but the pattern is very noisy. If we disregard the instances with high error and fit a linear regression to the points below 0.1, we find a slope of -4.97×10^{-6} . While this is negative, it is incredibly slight and it has a corresponding p -value of 0.70. Whatever downward trend may exist, it is difficult to detect.

For the next distribution we test, we set b in the second constraint for the distribution to be 1, forcing the derivative to be continuous at x_{\min} . The performance of each algorithm in determining

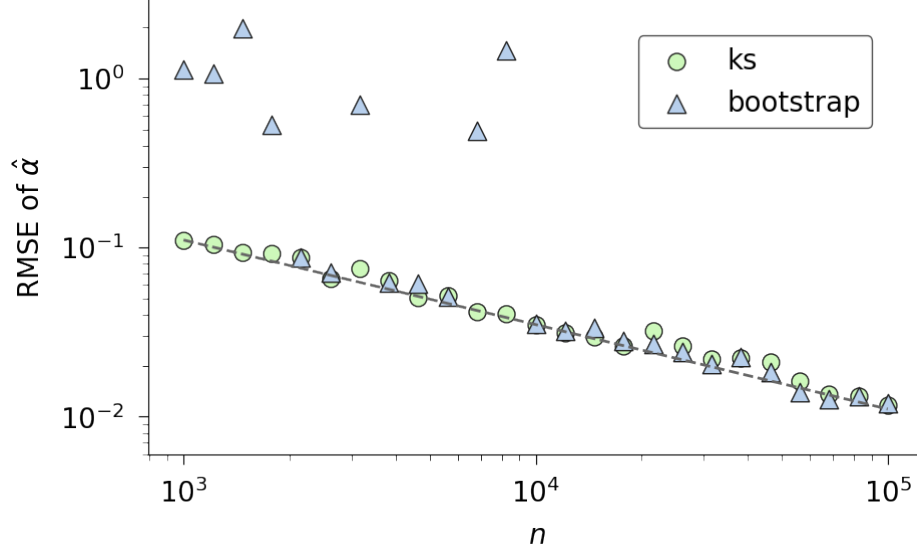


Figure 5.6: RMSE convergence comparison for the two methods on C^1 synthetic data sets by size with 500 bootstrap resamples. Each point is the RMSE over 250 synthetic data sets. The dashed line is the theoretical limit $n^{-1/2}$.

\hat{x}_{\min} changes slightly when this underlying distribution has a continuous derivative (Fig. 5.5). Consistent with expectations [32], both methods consistently underestimates the true value by a small amount. By examining the residual errors, it is clear the bootstrap method is again slightly more accurate, especially for larger values of x_{\min} , but this difference is more subtle than on the data with discontinuous derivative.

The convergence of the RMSE of $\hat{\alpha}$ on the continuously differentiable data sets is also less smooth and requires larger data sets to get errors that align with standard convergence rates (Fig. 5.6). The two algorithms perform quite comparably. As with the discontinuous version of this test distribution, the bootstrap method performs much worse for small data sets, but as n grows, each method converges at the expected rate of $n^{-1/2}$.

The RMSE converges more slowly as a function of the number of resamples for the bootstrap method on the C^1 data sets (Fig. 5.7). It does exhibit a decreasing error, but is noisy. A regression through the points with $\text{RMSE} < 0.1$ has a slope of -1.064×10^{-5} and intercept of 0.048 and a p -value of 0.503, so again the downward trend is barely detectable.

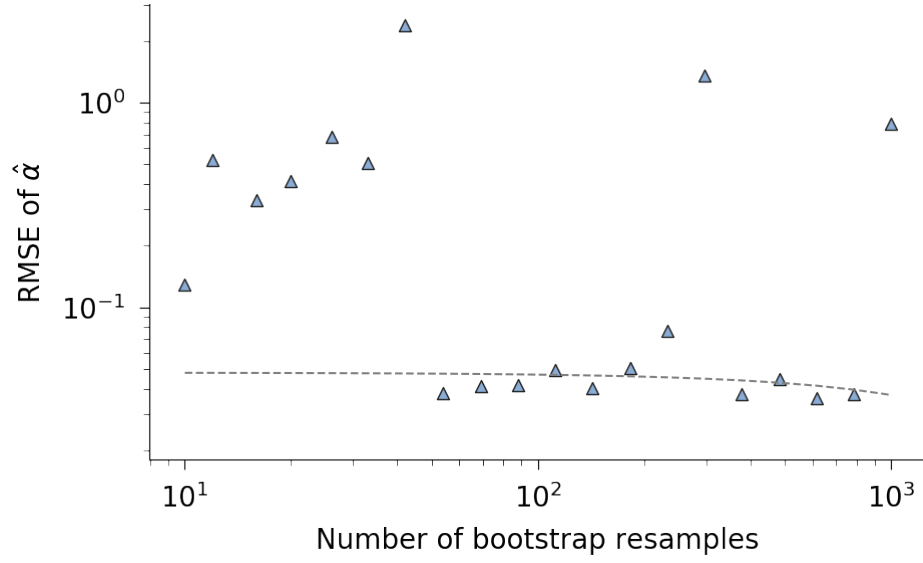


Figure 5.7: RMSE convergence comparison for bootstrap method on C^1 synthetic data sets by number of bootstrap resamples.

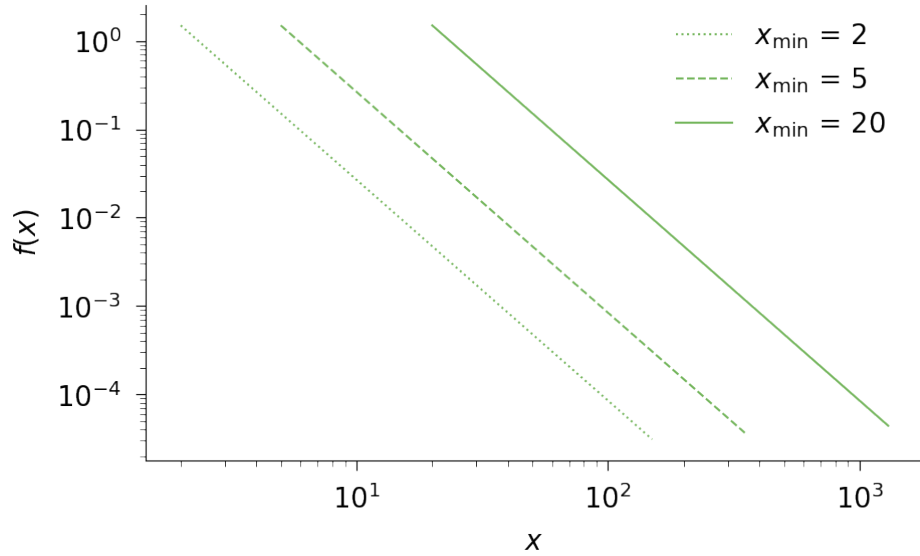


Figure 5.8: Pareto distribution

We also test the performance of the algorithms on samples drawn from a pure Pareto distribution, with a pdf of

$$f(x) = (\alpha - 1)cx^{-\alpha} \quad x \geq c^{\frac{1}{\alpha-1}}$$

where $c > 0$, $\alpha > 1$ (Fig. 5.8). The tail for these datasets is the entire span of x -values. The

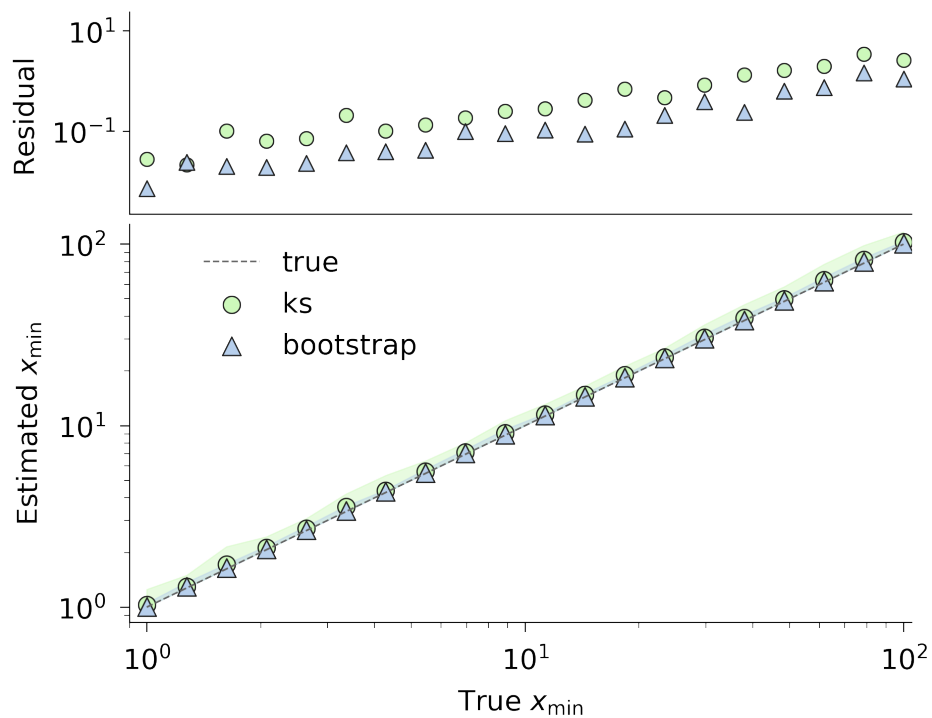


Figure 5.9: Recovery of x_{\min} for pareto synthetic data sets.

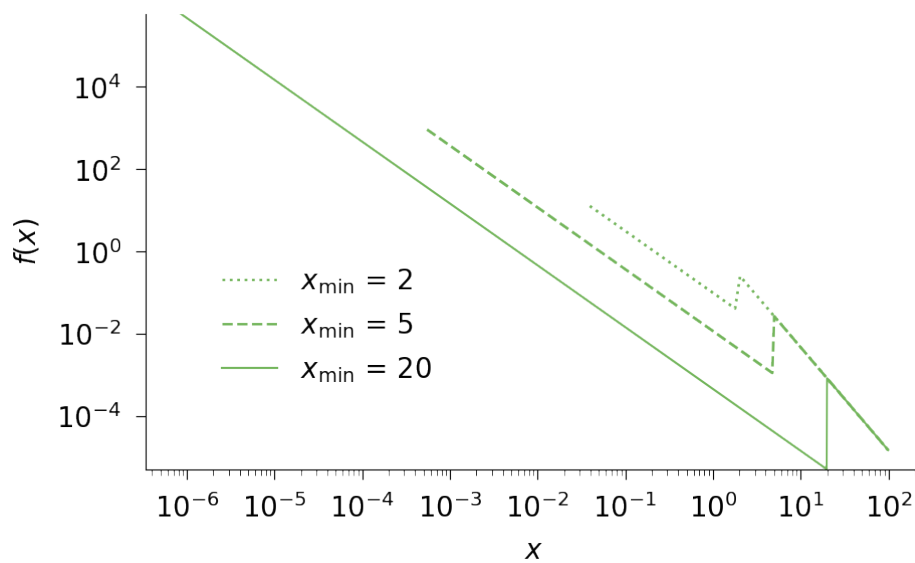


Figure 5.10: Pareto deviate distribution

KS method consistently slightly overestimates x_{\min} on these data sets (Fig. 5.9). The bootstrap method again exhibits slightly higher accuracy, which is more clearly seen in the residual error.

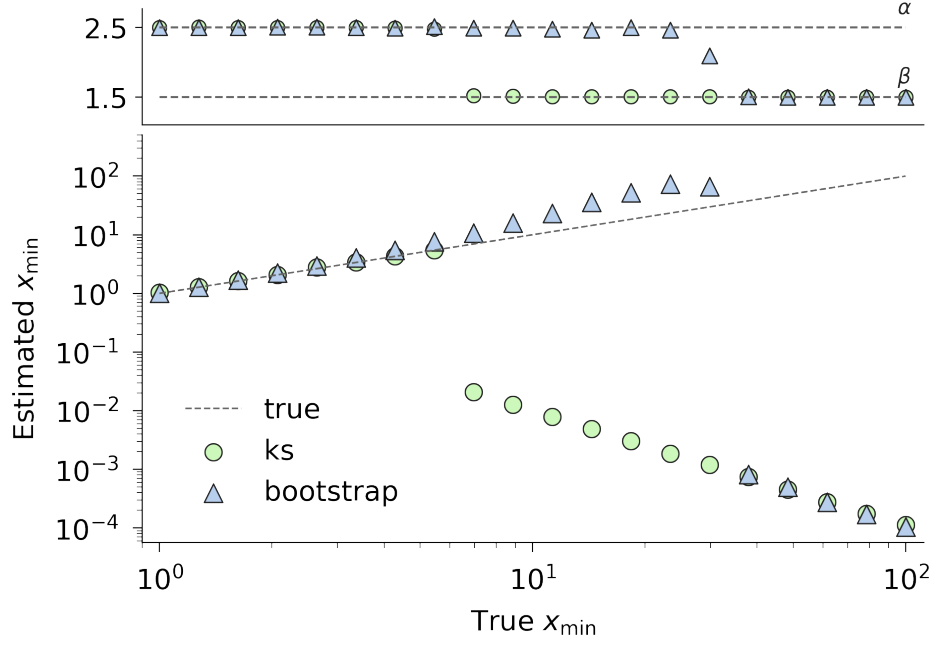


Figure 5.11: Recovery of x_{\min} for pareto deviate synthetic data sets. The estimate of α is shown above.

Distributions that deviate from a single Pareto can pose a particularly difficult problem [32]. Here we test a Pareto with exponent α above x_{\min} and Pareto with a different exponent $\beta \neq \alpha$ below x_{\min}

$$f(x) = \begin{cases} f_1(x) = (\beta - 1)t_0^{\frac{\alpha-\beta}{\alpha-1}}x^{-\beta} & \left(\frac{1}{t_0}\right)^{\frac{1}{\alpha-1} - \frac{1}{\beta-1}} \leq x < t_0^{-\frac{1}{\alpha-1}} \\ f_2(x) = (\alpha - 1)x^{-\alpha} & x \geq t_0^{-\frac{1}{\alpha-1}} \end{cases}$$

where $t_0 \in (0, 1)$ defines the probability of being in each piece of the function and $\alpha, \beta > 1$ (Fig. 5.10). Note that this pdf is discontinuous at x_{\min} .

The ability to recover x_{\min} on these data sets is different from the others (Fig. 5.11). Both methods are accurate for smaller values of x_{\min} , but as x_{\min} gets near and above about 8, the bootstrap method starts to overestimate x_{\min} , resulting in a smaller tail than the true tail, while the KS method quickly transitions to underestimating x_{\min} , and as the true tail of the data shrinks, KS method decreases x_{\min} estimates down towards zero.

Examining the estimate of α as a function of x_{\min} (Fig. 5.11) shows similar behavior, with

both methods recovering the true α with high accuracy for $x_{\min} < \sim 8$. Above this, the KS method estimates quickly converge on β , while the bootstrap estimates just slightly underestimate α . Above about $x_{\min} = 40$, the bootstrap method estimates of α align with the estimates from the KS method.

5.3.2 Real Data

We apply the methods to real data as well and compare the results. While we do not have ground truth to inform our assessment in this case, we note similarities and differences which might inform a choice between the two.

We analyze 22 real-world data sets, 13 with continuous data and 9 with discrete data, all of which have previously been studied for their potentially power-law behavior [22]. They cover a variety of different applications, ranging from the physical sciences to information sciences, engineering, and the social sciences. The continuous data sets as described in [22] are:

- (1) The numbers of sightings of birds of different species in the North American Breeding Bird Survey for 2003.
- (2) The numbers of customers affected in electrical blackouts in the United States between 1984 and 2002 [74].
- (3) The numbers of copies of bestselling books sold in the United States during the period 1895 to 1965 [42].
- (4) The human populations of U.S. cities in the 2000 U.S. Census.
- (5) The number of bytes of data received as the result of individual web (HTTP) requests from computer users at a large research laboratory during a 24-hour period in June 1996 [110]. Roughly speaking, this distribution represents the size distribution of web files transmitted over the Internet.
- (6) The sizes in acres of wildfires occurring on U.S. federal land between 1986 and 1996 [74].

- (7) Peak gamma-ray intensity of solar flares between 1980 and 1989 [74].
- (8) The intensities of earthquakes occurring in California between 1910 and 1992, measured as the maximum amplitude of motion during the quake [74].
- (9) The numbers of adherents of religious denominations, bodies, and sects, as compiled and published on the web site adherents.com
- (10) The frequencies of occurrence of U.S. family names in the 1990 U.S. Census.
- (11) The intensity of wars from 1816-1980 measured as the number of battle deaths per 10000 of the combined populations of the warring nations [98, 95].
- (12) The aggregate net worth in U.S. dollars of the richest individuals in the United States in October 2003 [74].
- (13) The number of “hits” received by web sites from customers of the America Online Internet service in a single day [1].

The discrete data sets are [22]:









- (1) The degrees of nodes in the partially known network representation of the Internet at the level of autonomous systems for May 2006 [47]. (An autonomous system is a group of IP addresses on the Internet among which routing is handled internally or “autonomously,” rather than using the Internet’s large-scale border gateway protocol routing mechanism.)
- (2) The number of citations received between publication and June 1997 by scientific papers published in 1981 and listed in the Science Citation Index [90]
- (3) The degrees of metabolites in the metabolic network of the bacterium *Escherichia coli* [49].
- (4) The sizes of email address books of computer users at a large university [75].
- (5) The number of species per genus of mammals. This data set, compiled by Smith et al. [99], is composed primarily of species alive today but also includes

some recently extinct species, where “recent” in this context means the last few tens of thousands of years.

- (6) The number of academic papers authored or coauthored by mathematicians listed in the American Mathematical Society’s MathSciNet database. (Data compiled by J. Grossman.)
- (7) The degrees (i.e., numbers of distinct interaction partners) of proteins in the partially known protein-interaction network of the yeast *Saccharomyces cerevisiae* [53].
- (8) The severity of terrorist attacks worldwide from February 1968 to June 2006, measured as the number of deaths directly resulting [26].
- (9) The frequency of occurrence of unique words in the novel *Moby Dick* by Herman Melville [74].

As in [22], we note that there may be biases in the way some of these data sets were sampled from larger systems. Our focus is on comparing the methods, so we are not concerned with the accuracy or interpretation of results for a specific data set. We therefore make no efforts to correct for any potential biases.

We focus first on the continuous data sets (Table 5.1). The KS method estimates of \hat{k} are almost always larger than the bootstrap estimates. This means the KS method is including more points in the power-law tail. This also means that the KS method estimates of x_{\min} are typically lower. The impact on $\hat{\alpha}$ is small: most estimates are similar between the two methods. The difference is most pronounced in the birds, fires, and flares data sets.

Data set	n	$\hat{\alpha}$		$\hat{\alpha}_{\min}$		\hat{k}		$\hat{\alpha}$ Distribution
		Boot	KS	Boot	KS	Boot	KS	
birds	204	4.011	2.139	83906	6679	2	65	
blackouts	199	2.258	2.27264	400000	230000	29	58	
books	633	3.75514	3.658	3.156×10^6	2.4×10^6	69	138	
cities	10088	2.398	2.370	75720	52457	359	579	
http	32630	2.462	2.483	35605	37116	7121	6793	
fires	7898	3.06369	2.164	96310	6324	16	520	
flares	5224	11.738	1.788	202600	323	2	1710	
quakes	13561	1.774	1.640	3162	794	5468	11696	





Data set	n	$\hat{\alpha}$		$\hat{\alpha}_{\min}$		\hat{k}		$\hat{\alpha}$ Distribution
		Boot	KS	Boot	KS	Boot	KS	
religions	103	1.7434	1.805	1.04394×10^6	3.85×10^6	101	38	
surnames	2753	2.174	2.493	12425.5	111919	2717	238	
wars	115	1.807	1.729	3.5	2.1	52	69	
web hits	119724	1.80513	2.090	1	42	68786	3429	

Table 5.1: Output of each method on continuous real data sets.

5.4 Discussion

Overall, both methods appear fairly comparable in output and accuracy in fitting the tail of a power-law distribution to a data set. On most data sets, the accuracy in estimates of both the threshold x_{\min} and the parameter α are high. We do see a slightly lower error in \hat{x}_{\min} for the bootstrap method, particularly as the true value of x_{\min} increases. In practice, this small increase in accuracy may or may not be important, depending on the application. On the first two data sets we find the convergence of the RMSE of $\hat{\alpha}$ looks close to the theoretical rate. Due to the randomness in the algorithm, the bootstrap method exhibits inconsistent and often much larger errors for data sets smaller than about $n = 10^4$. This is a practical concern because many data sets corresponding to complex systems, to which we often try to fit a power-law distribution, are smaller than this limit. For these data sets perhaps the KS method would be a better choice, to avoid the risk of having an error in α on the order of 1. However, for larger data sets, the RMSE is consistently slightly better for the bootstrap method than for the KS method. On larger data sets, then, the bootstrap method is a good choice. These observations are consistent with the theory that we have asymptotic convergence guarantees of the bootstrap method and lack such guarantees for the KS method.

The case of the piecewise Pareto distributions is unusual. The disagreement in the \hat{x}_{\min} values between the methods when x_{\min} takes intermediate values (meaning that the relative ‘size’ of the two power law regions is relatively close) seems to reflect the fundamental difference between the two approaches – both are trying to estimate the location of where the tail begins, but the KS method has a bias toward locations that increase the induced sample size. Both methods struggle when the true tail region is small compared to a statistically plausible (meaning, also power-law distributed) but relatively large non-tail region. This is the only case among the four synthetic distributions where we find any parameterization that leads to strongly different behavior by the two methods. Hence in most circumstances we can trust our conclusions above: when data sets are large, we should likely expect that they would produce similar results and that the bootstrap

method has higher accuracy, while when data sets are smaller, we might expect the KS method to give more reliable results.

It is less clear how to compare performance of the two methods when they are applied to real data. We observe that most estimates of α are similar, even when the estimates of x_{\min} appear different. Since our goal is often to prioritize accuracy in α , we can consider these data sets to yield similar algorithmic performance. In general, however, the trend appears to be smaller estimates of x_{\min} for the KS method than the bootstrap method. This means the KS method has more points in the tail. This has the potential to bias the estimate in α by fitting non-power-law points if the true x_{\min} is higher than the estimate. However, it is also possible that the bootstrap method is overestimating x_{\min} . The three data sets for which we have very different results for each method are the birds, fires, and flares data sets, all of which are smaller than $n = 10^4$. We have observed that the bootstrap method sometimes struggles on smaller data sets, so it is likely that this is why we see such different behavior here. The bootstrap method estimates very small values of k , and therefore does not have enough points in the tail to get a reliable estimate of α .

More study is necessary to fully understand the intricacies of the performance of the two methods, but we see some trends emerging already. The KS method is strong, particularly on moderately sized data sets with $10^3 - 10^4$ data points, but it lacks asymptotic convergence guarantees. The bootstrap method has asymptotic guarantees and performs quite well on larger data sets, but is sometimes unreliable on smaller data sets with fewer than about 10^4 data points. While the study here deals only with continuous data, it seems likely that the conclusions and observations would be similar in a discrete regime. In the size range of the degree sequences we analyze in Chapters 3 and 4, the KS method is a reasonable choice. Had we access to larger data sets for future analysis, we could consider analyzing those with the bootstrap method.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The idea of scale-free networks is prevalent in the literature. It is important to know when the properties of a network will theoretically follow known patterns. If a network has structure that is very different from scale free, it will behave differently. Researchers use modeling to understand dynamics in real-world systems. If the model makes assumptions that do not accurately represent the system, it is possible to reach conclusions about the dynamics that are not realistic. To prevent this from happening, we would like to do as well as possible at selecting models that are best for our system of study.

The methods and results we present in this thesis contribute to understanding aspects of model fitting and selection related to power laws. To select the ideal model for a system, it is crucial to thoroughly test the fit of models before selecting the best choice and to have fitting methods that are reliable. Chapter 3 presents methods for searching for power-law structure in degree sequences of any real-world networks. The results here also indicate that scale-free models may not be the best fit for all real-world networks. Scale-free models are theoretically very interesting and describe many systems quite well, but our findings suggest it is important to confirm for a particular network that these models are a good choice before proceeding with analysis.

Further, the exploration and tests we present in Chapter 4 suggest that the observations we made in the previous chapter are qualitatively robust to changes in methodology. This further emphasizes the importance of careful model selection. The goal of these chapters is not to determine

exactly which networks can and cannot be reasonably classified as scale free, but rather to shed light on the complexity of the problem and the value of model selection.

To address the choice of a specific model for a particular system or problem, we need to look more closely at the methodology we use to determine model fit. Chapter 5 compares the performance of two algorithms for finding the beginning of a power-law tail in empirical data. Overall we find both methods perform equivalently; sometimes KS had slightly higher bias and variance, while the bootstrap method had some convergence issues for alpha for moderate or small-sized data sets. This exploration also indicates that there can be nuances present in the data that make it particularly difficult for a certain algorithm, or sometimes for any algorithm, to be able to identify x_{\min} with reasonable accuracy. Further analysis is needed to assess how to determine the best method for this fitting, or if a universal best method exists at all. Additionally, it would be useful to develop methods for assessing the accuracy of the estimates for a real-world data set, e.g., by developing diagnostics for convergence or stability.

6.2 Future Work

The scale-free analysis in this thesis focuses exclusively on definitions of scale free in which the degree distribution of a network is in some sense a power law. This addresses a wide body of literature and common ideas in the network science community, but there are other notions of scale free we do not address. Another common idea in the literature is that a network is scale free if it was generated by some form of the preferential attachment mechanism. While this method is known to frequently generate networks with power-law degree sequences, this is not a guarantee and finding the generating mechanism for a network requires very different methodology from the techniques we use. Further study could examine whether this notion of scale-free networks is a better description than the power-law definition.

More subtly, there exist different ideas about what exactly is meant by the claim that a degree sequence follows a power-law. In extreme-value theory, instead of a power-law distribution

of the form

$$f(x) = Cx^{-\alpha} \quad x \geq x_{\min}$$

researchers consider *regularly-varying functions* of the form

$$f(x) = L(x)x^{-\alpha} \quad x \geq x_{\min}$$

where $L(x)$ is a *slowly-varying function*. A slowly-varying function satisfies

$$\lim_{x \rightarrow \infty} \frac{L(ax)}{L(x)} = 1$$

for all $a > 0$. Conceptually these functions change very slowly as $x \rightarrow \infty$. Thus the regularly-varying function, where we multiply a slowly-varying function by $x^{-\alpha}$ is asymptotically similar to the standard power law, but allows for deviations from strict power-law behavior at lower values of x . A recent study [106] used this framework to perform analysis similar to our work in Chapter 3. Through the lens of extreme-value theory, they found results that are slightly more inclusive than ours, because allowing deviations from the pure power-law distribution affords better fits of the distributions to the degree sequences. This underlines the importance of clearly defining what it means to call a network scale free. Future work could expand on this study with more rigorously tested methods for fitting discrete data and additional goodness-of-fit tests.

Study of methods for discrete data would also extend on the analysis in Chapter 5. The bootstrap method is proven to minimize the AMSE of the estimate of α for continuous data. Similar theory about the two methods we use or some alternative methods when applied to discrete data would be interesting and useful. Currently, the discrete version of the bootstrap method involves simply adding noise to the data before applying the continuous algorithm. Insight into methods that are specifically designed for discrete data might perform better or be easier to analyze. Additionally, the guarantee of asymptotic convergence does not necessarily mean that a method performs well on smaller data sets. Deeper study of methods in this regime would have far-reaching practical applications.

With additional understanding of discrete methods, we could perform similar analysis to what we present for synthetic continuous data, comparing the algorithms on discrete data from

different types of distributions. This would be useful for understanding what methods might work best on degree data. Here, future work could also explore methods for selecting which algorithm to use for real data. We need a way to compare performance the algorithms without access to ground truth.

Of course likelihood-based comparison methods are common, but this is an unusual case where they do not obviously apply. The bootstrap method is not a likelihood-maximization algorithm, so methods that compare likelihoods seem likely to favor the MSDP, biasing the comparison. Even without this issue, the likelihoods are not comparable when they select different values for x_{\min} ; whichever method chooses a smaller value will have more points in the tail of the distribution and therefore a smaller likelihood. Realistically, since the likelihood functions are identical for the two methods, this comparison ultimately does not make sense. Future work could study non-likelihood-based methods for model comparison to help find ways to assess which model is best for a given data set.

Bibliography

- [1] Lada A. Adamic and Bernardo A. Huberman. Technical Comment on Power-Law Distribution of the World Wide Web by A.-L. Barabási and R. Albert and H. Jeong and G. Bianconi. *Science*, 287(5461):2115a, 2000.
- [2] M. T. Agler, J. Ruhe, S. Kroll, C. Morhenn, S. T. Kim, D. Weigel, and E. M. Kemen. Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLoS Biology*, 14(1):1–31, 2016.
- [3] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for power law graphs. *Exp. Math.*, 10(1):53–66, 2001.
- [4] William Aiello, Fan R. K. Chung, and Linyuan Lu. A random graph model for massive graphs. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, May 21-23, 2000, Portland, OR, USA*, pages 171–180, 2000.
- [5] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.
- [6] R. Albert, H. Jeong, and A. L. Barabási. Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, 1999.
- [7] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.
- [8] D. L. Alderson and L. Li. Diversity of graphs with highly variable connectivity. *Phys. Rev. E*, 75:046102, 2007.
- [9] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA*, 97(21):11149–11152, 2000.
- [10] M. Ángeles Serrano Dmitri Krioukov and Marián Boguñá. Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.*, 100:078701, 2008.
- [11] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(October):509–512, 1999.
- [12] A.L. Barabasi. *Network Science*. Cambridge University Press, 2016.
- [13] A. Barrat, M. Barthélemy, and Vespignani A. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

- [14] N. Berger, C. Borgs, J. T. Chayes, R. M. D’Souza, and R. D. Kleinberg. Competition-induced preferential attachment. In *Proc. 31st International Colloquium on Automata, Languages and Programming (ICALP)*, pages 208–221, 2004.
- [15] B. Bollobás and O. Riordan. Coupling Scale-Free and Classical Random Graphs. *Internet Mathematics*, 1(2):215–225, 2004.
- [16] Anna D. Broido and Aaron Clauset. Scale-free networks are rare. *Nat. Commun.*, 10(1):1017, 2019.
- [17] Kenneth P. Burnham and David R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag, 2002.
- [18] G. Buzsáki and K. Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 15(4):264–278, 2014.
- [19] J. M. Carlson and J. Doyle. Highly optimized tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E*, 60:1412–1427, 1999.
- [20] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, England, 2008.
- [21] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.
- [22] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.
- [23] A. Clauset, E. Tucker, and M. Sainz. The Colorado Index of Complex Networks, `icon.colorado.edu`, 2016.
- [24] Aaron Clauset. Trends and fluctuations in the severity of interstate wars. *Sci. Adv.*, 4(2):26–28, 2018.
- [25] Aaron Clauset, Thuong T M Nguyen, Peter Hall, Tianchong Gao, Feng Li, Yu Chen, XuKai Zou, Bing-Rong Lin, Daniel Kifer, Jure Leskovec, Jon Kleinberg, Christos Faloutsos, M. P. H. Stumpf, and M. A. Porter. On Some Simple Estimates of an Exponent of Regular Variation. *Sci. Adv.*, 1(1):677, 2018.
- [26] Aaron Clauset, M. Young, and K.S. Gleditsch. On the frequency of severe terrorist events. *J. Conflict Resolut.*, 51:58–87, 2007.
- [27] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature Physics*, 2:110–115, 2006.
- [28] J Danielsson, L. De Haan, L Peng, and C. G. De Vries. Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. *J. Multivar. Anal.*, 76(2):226–248, 2001.
- [29] Laurens De Haan and Sidney Resnick. On asymptotic normality of the hill estimator. *Commun. Stat. Part C Stoch. Model.*, 14(4):849–866, 1998.
- [30] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51(September):1079–1187, 2002.

- [31] S. N. Dorogovtsev, J. F. F. Mendes, and a. N. Samukhin. Generic scale of the "scale-free" growing networks. *arXiv:cond-mat/0011115*, 2000.
- [32] Holger Drees, Sidney I Resnick, and Tiandong Wang. On a minimum distance procedure for threshold selection in tail analysis. *arXiv:1811.06433v1*, 2018.
- [33] T. DuBois, S. Eubank, and A. Srinivasans. The effect of random edge removal on network degree sequence. *Electronic Journal of Combinatorics*, 19:1–20, 2012.
- [34] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.
- [35] N. Eikmeier and D. F. Gleich. Revisiting Power-law Distributions in Spectra of Real World Networks. In *Proc. 23rd ACM SIGKDD Internat. Conference on Knowledge Discovery and Data Mining (KDD)*, pages 817–826, 2017.
- [36] Bailey K. Fosdick, Daniel B. Larremore, Joel Nishimura, and Johan Ugander. Configuring Random Graph Models with Fixed Degree Sequences. *arXiv preprint*, pages 1–42, 2016.
- [37] Bailey K. Fosdick, Daniel B. Larremore, Joel Nishimura, and Johan Ugander. Configuring Random Graph Models with Fixed Degree Sequences. *SIAM Rev.*, 2(60):315–355, 2017.
- [38] D. Gamermann, J. Triana, and R. Jaime. A comprehensive statistical study of metabolic and protein-protein interaction network properties. *arXiv:1712.07683*, 2017.
- [39] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:7821–7826, 2002.
- [40] K.-I. Goh, E. Oh, Hawoong Jeong, B. Kahng, and D. Kim. Classification of scale-free networks. *Proc. Natl. Acad. Sci. USA*, 99(20):12583–12588, 2002.
- [41] M. Golosovsky. Power-law citation distributions are not scale-free. *Phys. Rev. E*, 032306(June):1–12, 2017.
- [42] A. P. Hackett. *70 Years of Best Sellers, 1895-1965*. R.R.Bowker Company, New York, 1967.
- [43] S. Louis Hakimi. A remark on the existence of finite graphs. *Journal of the Society for Industrial & Applied Mathematics*, 1(11):135–147, 1963.
- [44] Peter Hall. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivar. Anal.*, 32(2):177–203, 1990.
- [45] Václav J. Havel. A remark on the existence of finite graphs. *Casopis Pest. Mat.*, 1253(80):477–480, 1955.
- [46] Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Stat.*, 3(5):1163–1174, 1975.
- [47] Petter Holme, Josh Karlin, and Stephanie Forrest. Radial structure of the Internet. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 463(2081):1231–1246, 2007.
- [48] Thomas House, Jonathan M. Read, Leon Danon, and Matthew J. Keeling. Testing the hypothesis of preferential attachment in social network formation. *EPJ Data Science*, 4(1):13, Oct 2015.

- [49] M. Huss and P. Holme. Currency and commodity metabolites: Their identification and relation to the modularity of metabolic networks. *IET Syst. Biol.*, 1:280–285, 2007.
- [50] Takashi Ichinomiya. Frequency synchronization in a random oscillator network. *Phys. Rev. E*, 70(2):5, 2004.
- [51] G. Ichinose and H. Sayama. Invasion of Cooperation in Scale-Free Networks: Accumulated versus Average Payoffs. *Artificial Life*, 23:25–33, 2017.
- [52] Kansuke Ikehara and Aaron Clauset. Characterizing the structural diversity of complex networks across domains. arXiv:1710.11304, 2017.
- [53] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA*, 97:1143–1147, 2000.
- [54] Matthew O. Jackson and Brian W. Rogers. Meeting strangers and friends of friends: How random are social networks? *Am. Econ. Rev.*, 97(3):890–915, 2007.
- [55] Hawoong Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.
- [56] R. Khanin and Ernst Wit. How scale-free are biological networks. *Journal of Comp. Bio.*, 13(3):810–818, 2006.
- [57] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models, and methods. *Computing and Combinatorics*, pages 1–17, 1999.
- [58] Deok Sun Lee. Synchronization transition in scale-free networks: Clusters of synchrony. *Phys. Rev. E*, 72(2):1–6, 2005.
- [59] S. H. Lee, M. D. Fricker, and M. A. Porter. Mesoscale analyses of fungal networks as an approach for quantifying phenotypic traits. *Journal of Complex Networks*, 5(1):145–159, 2017.
- [60] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution. *ACM Trans. Knowledge Discovery from Data*, 1(1):2–es, 2007.
- [61] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, and W. Willinger. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version). *Internet Math.*, 2(4):431–523, 2005.
- [62] Lun Li, David Alderson, Reiko Tanaka, John C. Doyle, and Walter Willinger. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version). arXiv: cond-mat/0501169, 2005.
- [63] G. Lima-Mendez and J. van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–1493, 2009.

- [64] Y. Malevergne, V. F. Pisarenko, and D. Sornette. Empirical Distributions of Log>Returns: between the Stretched Exponential and the Power Law? *Quantitative Finance*, 5(4):379–401, 2005.
- [65] David M. Mason. Laws of Large Numbers for sums of extreme values. *Ann. Probab.*, 10(3):754–764, 1982.
- [66] Deborah G. Mayo. *Error and the Growth of Experimental Knowledge (Science and Its Conceptual Foundations series)*. University of Chicago Press, 1996.
- [67] M. Middendorff, E. Ziv, and C. H. Wiggins. Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network. *Proc. Natl. Acad. Sci. USA*, 102(9):3192–3197, 2005.
- [68] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004.
- [69] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. 7th ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 29–42, 2007.
- [70] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2003.
- [71] M. Mitzenmacher. Editorial: The Future of Power Law Research. *Internet Mathematics*, 2(4):525–534, 2004.
- [72] T. Nakagawa and S. Osaki. The discrete weibull distribution. *IEEE Trans. Reliability*, 24:300–301, 1975.
- [73] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1), 2002.
- [74] M. E. J. Newman. Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46(5):323–351, 2005.
- [75] M. E. J. Newman, Stephanie Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66:035101, 2002.
- [76] M. E. J. Newman, M. Girvan, and J. D. Farmer. Optimal design, robustness, and risk aversion. *Phys. Rev. Lett.*, 89(2):028301, 2002.
- [77] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, 2003.
- [78] M.E.J. Newman. *Networks: An Introduction*. Oxford Univerity Press, 2010.
- [79] Newman, M. E. J. The first-mover advantage in scientific publication. *EPL*, 86(6):68001, 2009.
- [80] A. Pachon, L. Sacerdote, and S. Yang. Scale-free behavior of networks with the copresence of preferential and uniform attachment rules. arXiv:1704.08597, 2017.

- [81] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, 2015.
- [82] R. Pastor-Satorras, E. Smith, and R. V. Solé. Evolving protein interaction networks through gene duplication. *Journal of Theor. Biol.*, 222:199–210, 2003.
- [83] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, Apr 2001.
- [84] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics in finite size scale-free networks. *Phys. Rev. E*, 65:035108, Mar 2002.
- [85] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics in finite size scale-free networks. *Phys. Rev. E*, 65(3):1–4, 2002.
- [86] D. J. de S. Price. Networks of Scientific Papers. *Science*, 149(3683):510–515, 1965.
- [87] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):177–183, 2007.
- [88] Yongcheng Qi. Bootstrap and empirical likelihood methods in extremes. *Extremes*, 11(1):81–97, 2008.
- [89] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward. *P. N. A. S.*, pages 1–5, 2008.
- [90] S. Redner. How Popular is Your Paper? An Empirical Study of the Citation Distribution. *The European Physical Journal B*, 134:131–134, 1998.
- [91] S. Redner. Citation Statistics from 110 Years of Physical Review. *Phys. Today*, 58(June):49–54, 2005.
- [92] Juan G. Restrepo, Edward Ott, and Brian R. Hunt. Onset of synchronization in large networks of coupled oscillators. *Phys. Rev. E*, 71(3):1–12, 2005.
- [93] Juan G. Restrepo, Edward Ott, and Brian R. Hunt. Emergence of synchronization in complex networks of interacting dynamical systems. *Phys. D*, 224(1-2):114–122, 2006.
- [94] Juan G Restrepo, Edward Ott, and Brian R Hunt. Synchronization in large directed networks of coupled phase oscillators. *Chaos*, 16(1):015107, 2006.
- [95] D. C. Roberts and D. L. Tucotte. Fractality and self-organized criticality of wars. *Fractals*, 6(4):351–357, 1998.
- [96] C. Seshadhri, A. Pinar, and Tamara G. Kolda. An In-Depth Analysis of Stochastic Kronecker Graphs. *Journal of the ACM*, 60(2):1–30, 2011.
- [97] H. A Simon. On a Class of Skew Distribution Functions. *Biometrika*, 42:425–440, 1955.
- [98] M. Small and J. D. Singer. *Resort to Arms: International and Civil Wars, 1816–1980*. Sage Publications, Beverley Hills, CA, 1982.

- [99] F. A. Smith, S. K. Lyons, S. K. M. Ernest, K. E. Jones, D. M. Kaufman, T. Dayan, P. A. Marquet, J. H. Brown, and J. P. Haskell. Body mass of late quaternary mammals. *Ecology*, 84:p. 3403., 2003.
- [100] C. Song, S. Havlin, and H.A. Makse. Self-similarity of complex networks. *Nature*, 433:392, 2005.
- [101] M. P. H. Stumpf and M. A. Porter. Critical Truths About Power Laws. *Science*, 335(6069):665–666, 2012.
- [102] Michael P. H. Stumpf, Carsten Wiuf, and Robert M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221–4224, 2005.
- [103] R. Tanaka. Scale-rich metabolic networks. *Phys. Rev. Lett.*, 94(16):1–4, 2005.
- [104] Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.*, 8:32, dec 2012.
- [105] I. Voitalov. Tail index estimation for degree sequences of complex networks, 2018. <https://github.com/ivanvoitalov/tail-estimation>.
- [106] Ivan Voitalov, Pim Van Der Hoorn, Remco Van Der Hofstad, and Dmitri Krioukov. Scale-free Networks Well Done. 2018.
- [107] Q. H. Vuong. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333, 1989.
- [108] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
- [109] W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the Internet: A Source of Enormous Confusion and Great Potential. *Notices of the AMS*, 56(5):586–599, 2009.
- [110] W Willinger and V Paxson. Where Mathematics meets the Internet. *Not. Am. Math. Soc.*, 45(8):961–970, 1998.
- [111] L. Zhang, M. Small, and K. Judd. Exactly scale-free scale-free networks. *Physica A*, 433:182–197, 2015.